Master Économétrie et Statistiques, parcours Économétrie Appliquée

Mémoire de Master 1

L'outil 'Google Trends' peut-il aider à prévoir le chômage ?

Étude du taux de chômage français de 2004 à nos jours

THIERRY Diane

Sous la direction de Mr O.Darné

Juin 2020

Sommaire

Résumé	**2**
Introduction	**3**
Partie 1 : Environnement économique	**7**
Partie 2 : Méthodologie économétrique	**23**
Partie 3 : Présentation des données et application	**36**
Conclusion & Discussion	**74**
Bibliographie	**76**
Annexes	**80**
Table des matières	**86**

Résumé

Dans cette analyse nous cherchons à voir si l'utilisation des données de recherche Google est utile pour prévoir le taux de chômage en France par rapport à d'autres variables prédictives plus traditionnelles. Nous étudions ainsi le taux de chômage de 2004 à nos jours en prenant des variables classiques et une variable issue de l'outil Google Trends. Nous utilisons la méthodologie Box-Jenkins pour définir un modèle benchmark sur notre Y_t qu'est le taux de chômage ; nous trouvons ainsi un ARIMA[1,1,0] qui est à la fois pertinent et valide, et nous permet de réaliser des prévisions sur 3 horizons : h=1, h=4 et h=8. Par la suite nous cherchons à cointégrer les variables explicatives de manière à n'inclure dans des modèles ARX que celles qui ont une vraie relation stable avec Y_t. Elle est concluante pour deux d'entre elles ; la popularité du mot 'emploi' et la production industrielle. Après avoir construit des modèles et réalisé des prévisions à partir de ces derniers, nous avons vu que la variable Google Trends permet d'améliorer significativement les prévisions du taux de chômage. En effet, lorsqu'elle est ajoutée aux modèles ARX, la précision des prévisions qui y sont issues est beaucoup plus importante que celle des modèles composés seulement de variables dites "classiques". Son apport est d'autant plus important à court et moyen terme.

Mots-clefs : Google Trends, chômage français, cointégration, modèles ARX, prévisions.

Introduction

L'ère de la technologie s'est installée depuis un demi-siècle déjà sur la Terre. Nos modes de vie, nos habitudes, nos goûts ont été révolutionnés par cette transformation numérique qui vient changer notre société en profondeur. Le monde connecté dans lequel nous vivons aujourd'hui est caractérisé par le développement de l'intelligence artificielle, la mise en commun des ressources grâce aux réseaux et la connexion constante aux technologies. Nous vivons à présent dans un monde où tout est intelligent : les voitures peuvent avancer et se diriger seules, certaines lunettes permettent de faire des recherches en temps réel sur internet par un simple clin d'œil, des robots s'apparentent tellement aux humains qu'il devient difficile de les distinguer. Tout cela fascine certains, effraie d'autres et fait naître en eux des sentiments d'insécurité face aux avancées technologiques.

Ainsi, l'époque industrielle a fait place à une nouvelle génération de données et d'applications où chaque jour environ 2,5 trillions d'octets de données sont créés. Ce phénomène de Big Data, c'est-à-dire le stockage de milliers d'informations sur une base numérique, est apparu dès 1997 et ouvre aujourd'hui un champ presque infini de possibilité d'analyses et de recherches.[1]

De cette manière, depuis quelques années, certains économistes se concentrent sur le traitement de telles données au travers d'enquêtes conjoncturistes notamment, dont l'objectif selon l'INSEE est de "suivre la situation économique du moment et de prévoir les évolutions à court terme".[2] En ce sens, de nombreux outils de mesure informatiques ont fait leur apparition tels que "Google Trends", "Buzzsumo", "Reddit", "Social Share" etc., révélant les tendances de recherche des usagers. Ils permettent ainsi de visualiser le comportement des consommateurs en temps réel : les tendances d'un terme ou d'un autre reflètent les besoins et les désirs des utilisateurs d'internet. Ces données constituent donc une source d'information essentielle dans l'amélioration et la précision des prévisions économiques. Dans ce dossier nous nous pencherons plus particulièrement sur ce premier outil mis en place par le géant du Web 'Google'.

Aussi appelé 'Google Tendances de recherches' ou simplement 'Google Tendances', l'outil a fait son apparition en 2006, la même année que le rachat de la plateforme

[1] https://www.lebigdata.fr/definition-big-data (consulté le 20/01/2020)
[2] https://www.insee.fr/fr/metadonnees/definition/c1422 (consulté le 19/01/2020)

mondialement connue “Youtube” par Google. Ce dernier a pour but d’identifier la popularité de certains termes dans les recherches internet des individus à une période donnée. Ainsi, les valeurs des fréquences de recherche sont calibrées entre 0 et 100, de manière à faciliter la lecture et l'interprétation. Pour cela 2 cas possibles ; si l’on cherche l’évolution de la recherche d’un mot dans le temps, alors les données seront étalonnées par rapport au taux d’utilisation le plus élevé qui prendra la valeur 100. En revanche si l’on souhaite comparer les recherches d’un même mot sur différentes zones géographiques, Google Trends relativise le nombre de recherches de ce terme en particulier dans une zone, par rapport au nombre total de recherches effectuées sur le moteur ‘Google’ sur cette même zone - ce qui permet de comparer intelligemment différents pays ou régions. Ces données sont mises à jour quotidiennement dans le monde entier et dans toutes les langues gérées par Google. De plus, elles sont agrémentées de graphiques d'évolution dans le temps, ou de cartes animées de fréquences de recherche par région, par ville et par pays.[3]

Lors de recherches sur un tel outil il est possible de saisir soit un “terme de recherche” soit un “sujet”, la différence est subtile mais intéressante à connaître puisque les résultats seront différents. Un ‘terme de recherche’ est un mot-clef dont la recherche englobe tous les autres mots-clefs contenant ce terme dans la langue de recherche - ainsi, en recherchant le mot “mathématiques” les résultats affichés prendront en compte “cours de mathématiques”, “énigmes mathématiques” etc. Le ‘sujet’, quant à lui, va inclure tous les autres termes qui correspondent au même concept, dans toutes les langues. Donc en cherchant le même mot, cette fois les résultats incluront aussi “épreuves baccalauréat” ou “spécialité nouvelles filières” par exemple.[4]

Comme expliqué précédemment, les données mises à disposition par Google Tendances ne sont pas brutes, elles sont ajustées pour faciliter la comparaison entre les termes - l’outil ne donne donc pas le **volume des recherches** (quantité) mais bien la **popularité relative** d’un terme (intérêt) par rapport au nombre total de recherches effectuées dans la même zone géographique et à la même période. Par conséquent, l’outil Google Trends offre la possibilité de connaître en temps réel les tendances de recherches qui traduisent à la fois les goûts, les envies, mais aussi les préférences des internautes, représentant une mine d’informations très précieuse pour les conjoncturistes ainsi que pour

[3] https://www.abondance.com/20090819-10009-google-insights-for-search-disponible-en-francais.html (consulté le 15/01/2020)
[4] https://www.latranchee.com/comment-passer-de-lidee-a-la-strategie-grace-a-google-trends/ (consulté le 15/01/2020)

les professionnels du marketing et de la publicité. La gratuité et l'accessibilité ont rendu l'outil d'autant plus incontournable dans l'analyse des séries temporelles, grâce à la disponibilité de l'information en *open data* qui facilite les recherches des utilisateurs. En 2019 en France il y avait 53,1 millions d'internautes, soit près de 85% de la population de 2 ans et plus.[5] Sachant que 95% des recherches internet sont effectuées sur le moteur 'Google' (tout appareil confondu),[6] il semble que les données fournies par Google Trends soient assez représentatives de la réalité.

Nous pouvons nous demander si cet outil puissant est réellement un bon indicateur des tendances du Web, et si son attractivité est à la hauteur de son efficacité. Sa praticité a suscité un intéressement croissant chez différents économistes qui ont évalué au travers d'analyses, si l'outil Google Trends peut aider à la prédiction de certains indicateurs économiques. Les différents travaux orientés sur l'emploi et le chômage ont été réalisés dans de nombreux pays, tels que la Turquie (F.Bolívar, A.Ortiz et T.Rodrigo en 2019), l'Espagne (M.González-Fernández et C.González-Velasco en 2018), la France (Y.Fondeur et F.Karamé en 2013), l'Italie (F.D'amuri et J.Marcucci en 2017) et les États-Unis (B.Maas en 2019, S.Nagao, F.Takeda et R.Tanaka en 2019, et d'autres encore). Les conclusions de ces analyses diffèrent selon les variables explicatives utilisées et selon l'horizon de prédiction. Il apparaît en effet que Google Trends soit plus utile à la prévision de court terme que de long terme (comme le suggère l'étude réalisée par B.Maas en 2019 aux USA), en revanche la précision des modèles où est ajoutée la variable de Google est toujours supérieure à celle des modèles avec seulement les variables dites traditionnelles.

Cependant, si nous pouvons faire une remarque sur ces nombreux travaux réalisés, il apparaît que les auteurs n'ont pas vérifié la nature de la relation qui lie chaque variable explicative à Y_t, avant de les inclure dans les différents modèles. Ils ont ainsi pu en ajouter alors que leur relation avec la variable à expliquer était fallacieuse. De plus, la stationnarité des variables n'a pas été vérifiée. Notre objectif dans cette étude est donc de venir compléter ces travaux, pour que l'analyse soit complète et effective.

[5] https://www.journaldunet.com/ebusiness/le-net/1071394-nombre-d-internautes-en-france/ (consulté le 19/01/2020)
[6] https://www.inwin.fr/blog/un-outil-puissant-et-gratuit-mais-meconnu-google-trends/ (consulté le 19/01/2020)

Ainsi, dans ce dossier nous analyserons 6 séries temporelles équidistantes, avec le taux de chômage comme variable à expliquer, que nous étudierons puis modéliserons par un processus ARIMA, pour finalement effectuer des prévisions à 3 horizons différents. Par la suite nous construirons des modèles ARX où nous n'intégrerons que les variables explicatives cointégrées à Y_t, modèles que nous confronterons à l'aide de plusieurs critères de comparaison.

Le mémoire s'organise de la manière suivante : la partie 1 met en place l'environnement économique de ce travail par la justification et la définition des variables que nous utiliserons dans cette analyse. Dans la partie 2 nous expliquons toute la méthodologie de l'analyse avec les tests, les fonctions et les étapes, que nous suivrons dans la partie 3 qu'est la présentation des données et l'application des méthodes économétriques.

Cette troisième partie nous permet de prouver la contribution de la variable Google Trends dans la prédiction du taux de chômage puisque son ajout aux modèles, améliore amplement la qualité des prévisions.

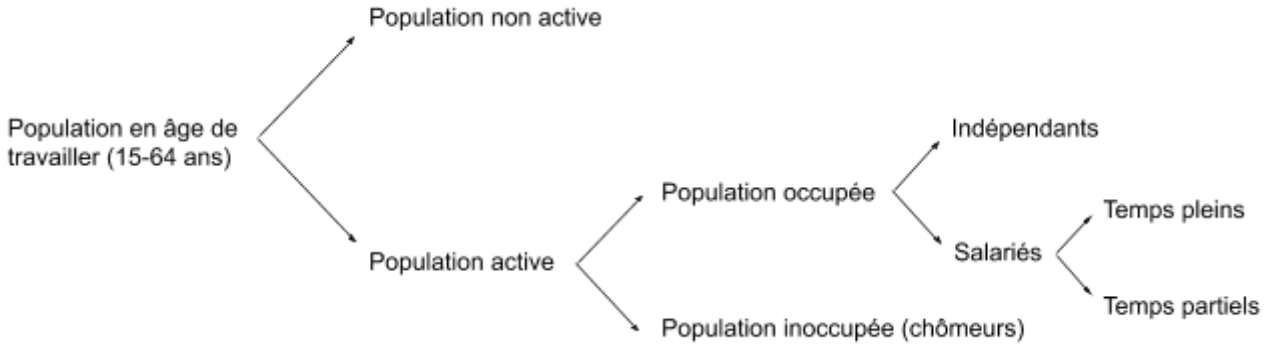
Partie 1 : Environnement économique

I. Analyse économique du sujet

L'année 2019 a été bénéfique d'un point de vue de l'emploi. En effet, avec 8.5% de chômage, la France a enregistré son plus faible taux depuis la crise de 2008 grâce à la création de 260.000 postes contre 188.000 l'année précédente.[7]

Ce dernier correspond selon l'INSEE au "pourcentage de chômeurs dans la population active" comme visible sur le schéma suivant, aussi un chômeur est une personne n'ayant pas de travail mais en cherchant activement un.

SCHÉMA N°1 : Découpage de la population en termes d'emploi



source : élaboration propre à partir de l'outil "dessin" sous google drive

Dans le monde entier le chômage n'est apparu qu'en 1973 avec le premier choc pétrolier, avant cette date chaque pays était en situation de plein-emploi voire de suremploi (plus de demande que d'offre de travail) ce qui conduisait certains pays à faire venir des travailleurs étrangers pour pallier au manque de main-d'œuvre, faisant monter légèrement le taux de chômage. Les années 70 et 80 ont été marquées par une hausse constante du chômage dans les pays d'Europe comme aux États-Unis ou au Japon, avec une hausse plus importante pour les pays d'Europe. Puis le chômage revient à des taux plus raisonnables à partir des années 90 mais reste plus élevé en Zone Euro qu'aux États-Unis ou au Royaume-Uni par exemple.[8] Enfin, c'est avec la crise des Subprimes que les taux explosent de nouveau et qu'une scission se crée entre les pays d'Europe du Sud tels que l'Espagne, la Grèce et le Portugal, qui

[7].https://www.ouest-france.fr/economie/emploi/chomage/le-taux-de-chomage-au-plus-bas-depuis-dix-ans-6671745 (consulté le 24/01/2020)

[8] https://www.researchgate.net/publication/322540093_Economie_du_Travail_Master_1_Dossier (consulté le 23/01/2020)

ont un taux de chômage très important par rapport aux pays d'Europe du Nord. Aujourd'hui, il existe une certaine dualité du marché de l'emploi où l'on retrouve d'un côté les *insiders* que sont les travailleurs qualifiés, rémunérés correctement avec une sécurité de l'emploi et des temps pleins, et d'un autre côté les *outsiders* qui n'ont pas ou peu de qualifications et sont à cause de cela mal rémunérés, ont des conditions de travail plus difficiles et des coûts de licenciement extrêmement bas, ce qui les place dans une situation précaire.

De manière générale le taux de chômage est une variable qui connaît beaucoup de fluctuations : les périodes de récession se traduisent souvent par une hausse des licenciements, donc du nombre de chômeurs, alors que les périodes d'expansion sont synonymes de création d'emplois et de baisse du taux de chômage. En revanche, généralement, celui-ci a tendance à baisser quand la population active diminue, ou quand il y a beaucoup d'emplois à temps partiel car cela implique un temps de travail moins important donc des facilités à embaucher et une réduction du niveau de chômage. [9]

Aussi, nous pouvons prendre l'exemple de la France et de l'Allemagne en 2018 pour comparer les niveaux de chômage en prenant en compte la structure de leur marché du travail (caractérisé par la démographie, les heures travaillées, les qualifications...) et son impact sur l'emploi en général.

TABLEAU N°1 : Comparaison de l'emploi en 2018 : France vs Allemagne

	France	**Allemagne**	**Ratios de comparaison**
Taux de chômage (% de la population active)	9,1	3,4	$\frac{9,1}{3,4} = 2,67$
Nombre d'heures travaillées (par an)	1 520	1 363	$\frac{1520}{1363} = 1,12$
Taux d'emplois à temps partiel (% de l'emploi)	14	22	$\frac{22}{14} = 1,57$
Taux de fécondité (nb d'enfants/femme)	1,87	1,46	$\frac{1,87}{1,46} = 1,28$

[9] https://www.researchgate.net/publication/322540093_Economie_du_Travail_Master_1_Dossier (consulté le 23/01/2020)

Le tableau n°1 ci-dessus, dont les données sont extraites du site internet de l'OCDE[10], nous montre que le taux de chômage français était 2,7 fois plus élevé que le chômage allemand avec un taux à 9,1% en 2018. Seulement, il intéressant de noter qu'en moyenne les Français travaillent davantage que les Allemands (157 heures de différence, soit 1,12 fois plus) et que le pourcentage d'emplois à temps partiel est largement supérieur en Allemagne : 22% contre 14% en France. Cela s'explique par l'existence d'une sorte de 'pression sociale' sur les femmes pour qu'elles restent dans leur foyer s'occuper des enfants, ce qui justifie le taux 1,57 fois supérieur au taux français. Cependant, les temps partiel conduisent parfois à la précarité et à la hausse des inégalités. Enfin, l'accroissement démographique est plus important en France puisque le nombre moyen d'enfants par femme est de 1,87 contre 1,46 en Allemagne ; il y a ainsi un phénomène de vieillissement de la population allemande qui conduit à une baisse de la population active réduisant le taux de chômage ($= \frac{chômeurs}{population\ active}$). Par conséquent, l'Allemagne a vu son niveau de chômage se réduire grâce aux nombreux emplois à temps partiel (impliquant un nombre d'heures de travail plus faible), et à une baisse de la population active.

Le taux de chômage est donc une variable qui doit s'étudier dans un contexte, c'est-à-dire en prenant en compte d'autres facteurs qui caractérisent le marché du travail et permettent de mieux comprendre le niveau de chômage. En outre, nous utiliserons les informations de la variable Google Trends mais aussi de 4 autres variables reflétant à la fois le marché du travail et l'activité économique française de 2004 à nos jours, afin de prédire au mieux le taux de chômage (Y_t).

[10] https://data.oecd.org/fr/ (consulté le 24/01/2020)

II. Analyse des variables explicatives

Le but de cette étude est de voir si l'utilisation de 'Google Trends' peut améliorer les prévisions du chômage et de l'emploi, c'est pourquoi nous prendrons en compte une variable explicative issue de cet outil, mais aussi des variables dites plus "classiques" qui, selon la littérature, sont de bons indicateurs du taux de chômage.

Dans cette partie nous justifierons l'utilisation des différentes variables explicatives, en montrant leur impact sur le taux de chômage, puis nous visualiserons graphiquement la relation qu'il existe entre chaque variable avec Y_t. Ainsi nous verrons s'il s'agit d'une corrélation positive qui implique une hausse du taux de chômage lorsque la variable étudiée augmente, ou d'une corrélation négative qui suppose une évolution opposée des 2 variables.

De surcroît, nous distinguerons volontairement 4 périodes sur notre échantillon qui s'étend de 2004 à 2019, puisque les données ont été très largement affectées par la crise mondiale de 2008, il se pourrait donc que les relations qui existent entre Y_t et X_t diffèrent d'une période de temps à une autre. Nous distinguons donc les phases suivantes :

- **2004-2006** : la période précrise caractérisée par une bonne santé économique où le chômage diminue considérablement
- **2007-2009** : autrement dit la période d'éclatement de la bulle immobilière aux États-Unis en décembre 2007, à laquelle suivirent l'effondrement du système financier puis la propagation de la crise à bon nombre d'économies nationales
- **2010-2015** : durant ces 5 années les effets de la Grande Récession se font largement sentir sur l'économie française ce qui se traduit, d'un point de vue de l'emploi, par une explosion du taux de chômage atteignant 10,48% au dernier trimestre de l'année 2014
- **2016-2019** : enfin, à partir de 2016 le taux revient à un niveau plus faible, et l'on retrouve, comme expliqué dans la section précédente, des taux de chômage aussi bas qu'avant 2008

A) Google Trends : X_1

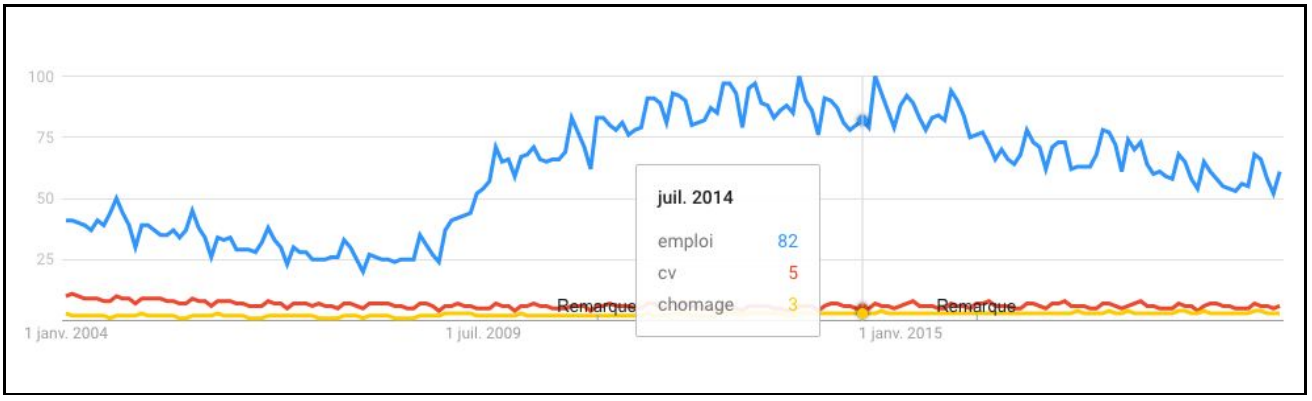
L'accès à internet n'a cessé de croître dans les foyers français depuis son apparition dans les années 70 tel qu'en 2018, près de 9 ménages sur 10 y avaient accès.[11] Ce système de

[11] https://fr.statista.com/statistiques/509227/menage-francais-acces-internet/ (consulté le 25/01/2020)

télécommunications informatiques permet à tout usager d'accéder à une immensité d'informations de différents types, à savoir articles, images, vidéos et d'autres encore. De plus en plus, les offres d'emploi se font en lignes : les demandeurs d'emploi (c'est-à-dire les entreprises) postent des annonces décrivant les emplois vacants et le profil type recherché, et les offreurs d'emploi (les ménages) accèdent à ces requêtes en ligne et peuvent ainsi consulter et répondre aux offres. Aujourd'hui, ce sont 88% des offreurs d'emploi qui utilisent internet pour effectuer leurs démarches de recherche. [12]

Ainsi, le développement d'outils permettant la visualisation des mots utilisés lors des recherches internet est particulièrement intéressant du point de vue de l'emploi. De plus, les données sur le chômage et l'emploi mettent 1 à 2 mois à sortir, en ce sens l'utilisation d'outils disponible instantanément peut être utile pour combler cet écart et prédire facilement le niveau d'emploi français. En outre, pour voir l'importance de Google Trends dans la prédiction du chômage il convient de choisir un mot s'y référant et regarder ses tendances de recherche (calibrées entre 0 et 100 comme expliqué en introduction). Pour choisir le mot adéquat nous comparons la popularité de recherches des mots "emploi", "CV" et "chômage" sur la période de 2004 à ce jour. Nous cherchons la popularité relative de chacun des mots entrés comme 'terme de recherche' et non comme 'sujet' dont la différence a été expliquée en introduction. Il est effectivement plus intéressant de saisir ici un terme de recherche puisque les résultats prendront en compte toute recherche effectuée avec ce mot : pour 'emploi' par exemple seront aussi incluses des recherches telles que "le pôle emploi", "le bon coin emploi", "offres d'emploi" etc.

GRAPHIQUE N°1 : Comparaison de la popularité des termes "emploi", "CV" et "chômage"



source : site internet de Google Trends

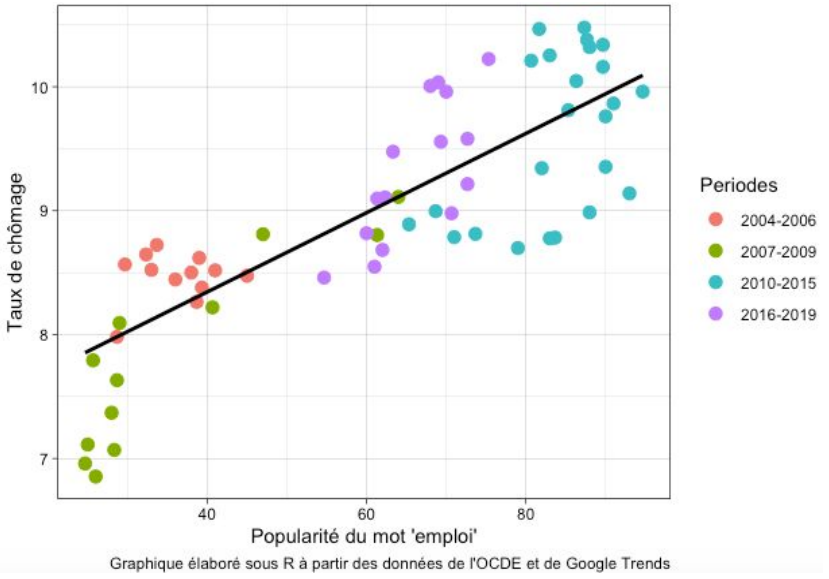
Le graphique n°1 nous donne visuellement les tendances de ces 3 mots, nous rappelons qu'il ne s'agit pas là des fréquences d'apparition mais bien de la popularité relative

[12].https://www.lefigaro.fr/emploi/2017/01/17/09005-20170117ARTFIG00290-les-francais-cherchent-un-emploi-sur-internet-mais-le-trouvent-grace-a-leur-reseau.php (consulté le 21/01/2020)

d'un terme par rapport au nombre total de recherches effectuées sur les mêmes périodes et régions comme c'est le cas ici. Nous constatons donc que le terme "emploi" est plus populaire que les 2 autres puisqu'il a les tendances de recherche les plus élevées sur la période étudiée, il semble ainsi mieux refléter les tendances de recherche des internautes en termes d'emploi. De plus, nous notons des pics réguliers dans la courbe d'évolution des recherches 'emploi', il s'agit en réalité des mois de septembre de chaque année où les recherches augmentent fortement et sont liées à la rentrée, il existe donc un phénomène de saisonnalité pour cette série. Nous décidons de retenir ce mot pour étudier l'apport de l'outil Google Trends pour prédire le taux de chômage, en ajoutant ses fréquences d'apparition comme variable explicative, de 2004 à ce jour.

Par conséquent, nous attendons une **relation positive** entre le taux de chômage et les recherches internet du mot 'emploi' ; en effet sa popularité dépend de l'intérêt suscité par cette recherche. En temps d'expansion où il n'y a généralement pas ou peu de chômage, les recherches d'emploi (que ce soit via des agences spécialisées ou via internet) sont faibles puisque le pays se trouve alors dans une situation proche du plein-emploi. En revanche lorsque l'activité économique ralentit les firmes se voient parfois contraintes de licencier leurs employés car la récession implique une baisse de la demande de la part des ménages dont le pouvoir d'achat diminue, qui conduit à une baisse de l'offre proposées par les entreprises comme réponse au ralentissement économique. Ainsi, l'intérêt des recherches liées à l'emploi augmente fortement et cela peut se mesurer notamment grâce à l'outil Google Trends.

<u>GRAPHIQUE N°2 : Relation entre le taux de chômage et la popularité du mot 'emploi'</u>



D'après le graphique de corrélation n°2 on observe effectivement une corrélation positive entre l'intérêt de la recherche du mot 'emploi' et le taux de chômage. Ainsi sur l'échantillon étudié, c'est-à-dire en France de 2004 à 2019, la théorie se vérifie. On peut noter que la grande majorité des observations se situent proches de la droite de corrélation, exceptés certains trimestres de 2007 à 2009 (couleur verte) où le taux de chômage, et donc les recherches Google Trends du mot 'emploi' sont faibles. Cette période précrise était en effet caractérisée par un chômage extrêmement bas dû, selon Christine Lagarde (ministre de l'Économie et de l'emploi à l'époque), à la création d'environ 340.000 emplois en 2007.[13]

On peut aussi noter, grâce au code couleur mettant en avant différentes périodes, que la relation positive entre le taux de chômage et la variable Google Trends existe par la combinaison de ces différentes périodes. En effet, si l'on s'intéresse à cette corrélation au sein des périodes, on voit qu'elle n'apparaît pas clairement positive par chacune d'entre elles. Par exemple, sur la période 2004-2006 (couleur rouge) on identifierait une relation légèrement négative, de même pour les trimestres entre 2010 et 2015 (couleur bleue) où on ne voit pas précisément de relation positive puisque les observations forment davantage un "bloc". Ce phénomène est connu sous le nom de **paradoxe de Simpson** (décrit à partir de 1951) ou effet de Yule-Simpson, et se produit lorsqu'un phénomène observé dans un ensemble diffère de celui observé dans les sous-ensembles qui le composent. Il n'est pas très marqué dans notre cas, mais c'est une chose à laquelle nous devons prêter attention pour que les conclusions de notre analyse soient bien établies.

B) Taux d'intérêt : X_2

Le taux d'intérêt représente le prix qu'il faut payer pour emprunter de l'argent, prix qui rémunère le service rendu par celui qui prête l'argent (il est exprimé en pourcentage).[14] Ainsi, les taux d'intérêt sont des déterminants importants de l'investissement des entreprises et de la consommation des ménages, principaux postes de la demande globale. En effet s'ils sont faibles, ces taux favorisent l'investissement des entreprises qui peuvent emprunter à des taux avantageux, au contraire ils freineront l'investissement s'ils sont élevés. Sachant que l'investissement détermine l'activité des entreprises donc la production globale et la croissance économique, il sera un bon indicateur du taux de chômage à travers l'offre

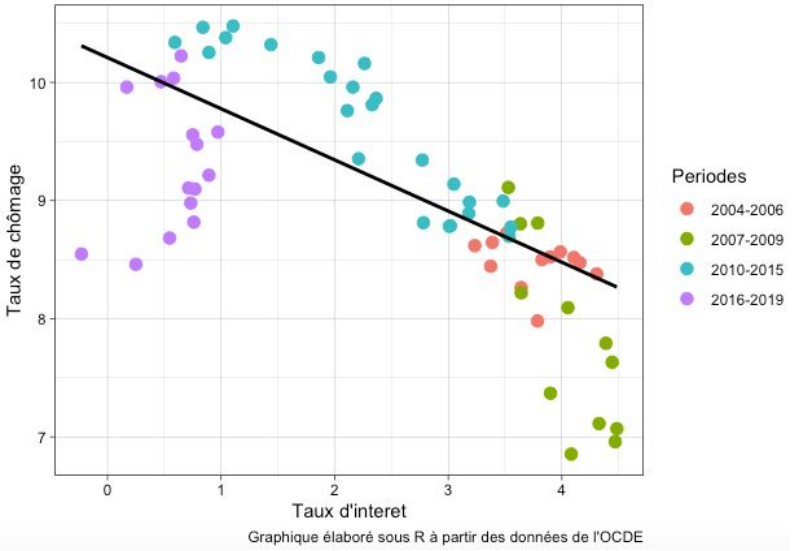
[13] https://lexpansion.lexpress.fr/actualite-economique/taux-de-chomage-a-8-en-2007_470864.html (consulté le 07/02/2020)
[14] https://www.insee.fr/fr/metadonnees/definition/c1287 (consulté le 07/02/2020)

d'emploi des firmes.[15] Il existe de nombreux taux d'intérêt à savoir ceux de court terme, de moyen/long terme, les taux réels et nominaux… La variable que nous avons choisie est celle du taux d'intérêt français à long terme. Il correspond aux obligations d'État à échéance de 10 ans, c'est-à-dire les cours auxquels ces obligations s'échangent sur les marchés de capitaux.[16]

Comme expliqué précédemment le taux d'intérêt peut favoriser l'investissement lorsqu'il est faible et le freiner s'il est élevé car les coûts d'emprunts sont alors trop importants pour les entreprises. S'il le favorise (taux d'intérêt faible) il permet aux entreprises d'augmenter leur production ce qui crée un climat économique favorable à la création d'emploi et à l'embauche, et a donc un effet positif sur la santé économique globale du pays, qui se traduit par une baisse du taux de chômage.[17] L'effet d'une hausse des taux d'intérêt sur le chômage peut passer par le canal de l'investissement des entreprises comme nous venons de le montrer, mais aussi par la consommation des ménages. En effet, de forts taux d'intérêt constituent une charge en plus pour les ménages diminuant alors leur propension à consommer, ce qui fait chuter la demande globale. En réponse à cela, les entreprises diminuent l'offre et donc leur production, augmentant ainsi le chômage parce que le besoin de main d'œuvre devient plus faible.

La relation théorique qui existe entre ces 2 variables est donc **positive** que ce soit via la FBCF ou via la demande globale, nous allons voir maintenant si sur notre échantillon cette relation est vérifiée.

GRAPHIQUE N°3 : Relation entre le taux de chômage et le taux d'intérêt



[15] https://data.oecd.org/fr/interest/taux-d-interet-a-long-terme.htm (consulté le 07/02/2020)
[16] https://data.oecd.org/fr/interest/taux-d-interet-a-long-terme.htm (consulté le 07/02/2020)
[17] https://www.persee.fr/doc/reco_0035-2764_1999_num_50_5_410125 (consulté le 17/02/2020)

Le lien qui apparaît sur le graphique n°3 est clair : il s'agit d'une relation négative contrairement à ce que l'on a pu supposer dans la partie économique. Cela peut être expliqué en partie par la particularité de la période étudiée, en effet celle-ci englobe la situation presque utopique d'avant crise (à savoir une production forte, un taux de chômage faible etc.) puis les effets de la crise mondiale qui est venue bouleverser l'équilibre économique des nations. On ne retrouve pas cette fois l'effet de Yule-Simpson puisque, même prises individuellement, les périodes sont caractérisées par des corrélations négatives entre Y_t et X_2. Seuls les trimestres compris entre 2016 et 2019 ne présentent pas la même relation entre les 2 variables ; en effet les taux d'intérêt négociés sont extrêmement bas avec un taux moyen négatif au troisième trimestre de 2019 (-0.23%). Cela est lié au taux directeur de la Banque Centrale Européenne (BCE) qui oriente les taux du marché et qui est volontairement institué à 0% (taux directeur nul) jusqu'au retour durable de l'inflation à 2%. Pour se faire, la BCE rachète de la dette publique et privée sur le marché, à hauteur de 20 milliards d'euros par mois de manière à réanimer l'activité économique via les prêts, l'investissement etc.[18] Ainsi depuis mars 2016 le taux directeur est nul (voir annexe n°1), ce qui explique les niveaux de taux faibles, ajoutés à un taux de chômage qui va diminuant, il apparaît que les points de 2019 sont inhabituels.

Mais pourquoi la relation entre Y_t et X_2 est radicalement contraire aux attentes fondées sur les théories économiques ? L'économiste américain Thomas Palley a tenté d'expliquer, grâce à une étude publiée en 2018,[19] comment des taux d'intérêt bas voire négatifs entraînent une hausse du chômage via 2 canaux différents. Premièrement, dans le cas où l'effet de revenu l'emporte sur l'effet de substitution, la baisse des taux d'intérêt conduit à l'augmentation de l'épargne car les ménages qui voient leurs flux futurs d'intérêt baisser, compensent ce manque à gagner en épargnant davantage. De plus, les taux négatifs sur les dépôts peuvent être perçus comme une taxe pour les ménages qui, par un sentiment de baisse de richesse, vont limiter leur consommation. La suite de ce processus est assez intuitive quant à la hausse du chômage via une baisse de la demande globale puis de la production des firmes. Dans un second temps, les taux d'intérêt négatifs peuvent non pas stimuler mais freiner l'investissement des entreprises, puisqu'elles passent par le canal bancaire pour obtenir des prêts leur permettant de financer de futurs investissements. Or, les banques voient leurs profits diminuer avec ces

[18] https://www.lefigaro.fr/conjoncture/la-bce-maintient-ses-taux-directeurs-au-plus-bas-1-20191212 (consulté le 07/02/2020)

[19] T.Palley, "*Negative Interest Rate Policy (NIRP) and the Fallacy of the Natural Rate of Interest: Why NIRP May Worsen Keynesian Unemployment*", PERI (Political Economy Research Institute), working paper, n° 463.

taux d'intérêt négatifs - elles seront donc moins enclines à prêter, et répercuteront leurs coûts de refinancement sur les taux appliqués aux emprunteurs provoquant ainsi une frilosité bancaire et un assèchement du crédit. Avec un manque d'investissement, les entreprises deviendront moins compétitives ce qui conduira à la hausse du chômage.[20] En conclusion, selon Palley, des taux d'intérêt faibles ou négatifs conduisent à un chômage élevé via une hausse de l'épargne de la part des ménages et une baisse de l'investissement de la part des entreprises, taux pourtant fixés à 0 en réponse à la crise de 2008. Cette théorie apporte une justification plausible de la relation empiriquement négative entre le taux de chômage et le taux d'intérêt.

C) Production industrielle : X_3

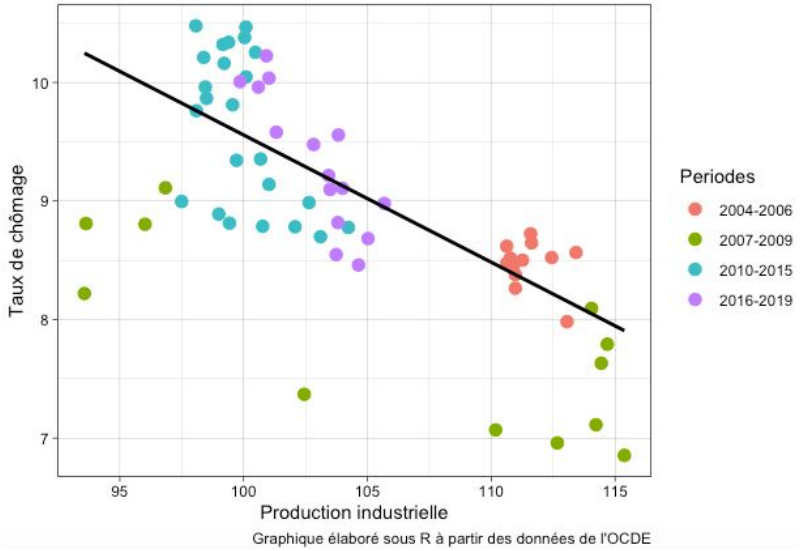
La production industrielle manufacturière désigne la production d'entités industrielles, elle englobe plusieurs secteurs d'activité tels que l'extraction minière, les activités manufacturières, l'électricité, gaz, eau et climatisation. Exprimée sous forme d'un indice base 2015=100 dans notre cas, la production industrielle reflète les variations du volume de production du secteur secondaire sur une période donnée.[21] Représentant 16,9% du PIB français en 2018,[22] la production industrielle est souvent utilisée comme indicateur de l'activité économique car elle suit de près ses variations. En effet, le secteur secondaire est fortement lié aux cycles économiques, ainsi, en phase d'expansion l'indicateur augmente et il diminue lors des périodes de récession. Les variations au sein du secteur secondaire sont souvent à l'origine des mouvements du PIB ; comme expliqué pour les taux d'intérêt, les périodes de croissance économique s'accompagnent de baisses du nombre de chômeurs, liées à la création de nombreux emplois pour répondre à une demande en hausse. A contrario, les périodes de ralentissement économique sont synonymes de licenciements et ainsi d'un accroissement du taux de chômage dans le pays.

En ce sens, la corrélation existante entre l'indice de production industrielle et le taux de chômage est **négative** : lorsque l'économie est en bonne santé (fort volume de production

[20].https://blogs.alternatives-economiques.fr/anota/2018/05/27/les-taux-d-interet-negatifs-peuvent-aggraver-le-chomage-keynesien (consulté le 15/05/2020)
[21] https://data.oecd.org/fr/industry/production-industrielle.htm (consulté le 07/02/2020)
[22] https://donnees.banquemondiale.org/indicateur/NV.IND.TOTL.ZS?view=chart (consulté le 07/02/2020)

industrielle donc hausse de l'indice) le taux de chômage a tendance à baisser. Voyons à présent si cette relation est vérifiée sur notre échantillon de données.

GRAPHIQUE N°4 : Relation entre le taux de chômage et la production industrielle



D'après le graphique n°4 on constate une corrélation négative entre ces 2 variables, ce qui confirme donc notre hypothèse. Une fois de plus les trimestres des années 2007, 2008 et 2009 sont excentrés du groupement des observations le long de la droite de corrélation ; l'indice chute à 93 points pour le premier trimestre de 2009. Les effets de la crise mondiale des Subprimes se ressentent donc largement sur la production nationale qui baisse progressivement au cours de l'année 2008 puis reprend de la vigueur et réaugmente jusqu'en 2019 (couleur bleue puis violette) où elle atteint 105 points.

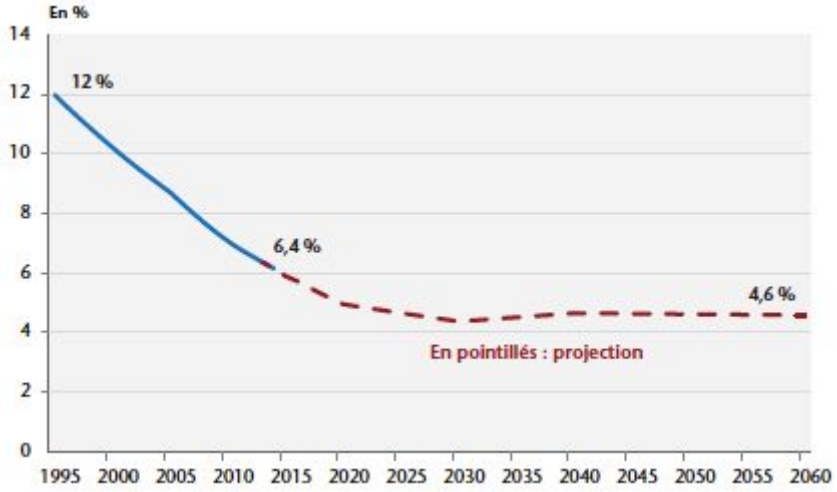
D) Population active : X_4

"Chaque année, en France, 800.000 jeunes entrent sur le marché du travail tandis que 670.000 seniors partent à la retraite. La population active progresse donc de 130.000 personnes, ce qui signifie qu'il faut créer de très nombreux postes pour que le chômage stagne ou ne progresse pas."[23] Cette explication d'Éric Heyer (docteur en Sciences Économiques), illustre le fait que l'accroissement de la population active entraîne une hausse de la population en âge de travailler, pour laquelle il est parfois difficile de trouver un emploi car les offres d'emploi deviennent supérieures aux demandes d'emploi. La population active correspond au nombre de personnes en âge de travailler c'est-à-dire les 15-65 ans actifs. C'est donc la population en âge de travailler à laquelle on retire les inactifs que sont les étudiants, les

[23] https://www.brief.eco/a/2018/10/31/on-fait-le-point/les-determinants-du-chomage/ (consulté le 23/01)

hommes/femmes au foyer et les retraités. Depuis quelques années, la France comme de nombreux autres pays, assiste à un vieillissement de sa population. Effectivement, le taux de fécondité français en 2018 était de 1,87 enfant par femme[24] alors que le taux qui permet de garder une population stable (sans vieillissement) est de 2,1.

GRAPHIQUE N°5 : Ratio "actifs âgés de 15-54 ans" sur "actifs de 55 ans et plus"



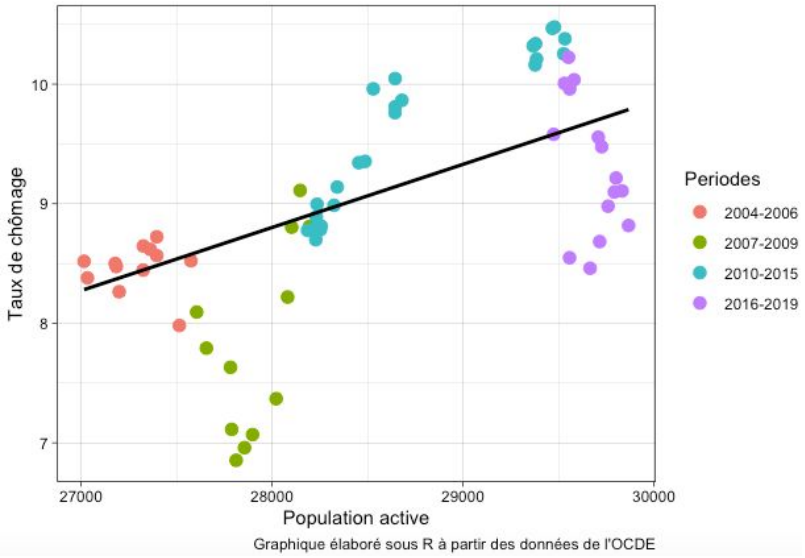
Source : https://www.cairn.info/revue-de-l-ofce-2015-6-page-339.htm (consulté le 09/02/2020)

Aussi, comme visible sur le graphique n°5, au cours des 30 dernières années la population active française a beaucoup vieilli ; en 1995 il y avait 12 fois plus d'actifs de 15 à 54 ans que d'actifs de 55 ans et plus - ce ratio tend à se stabiliser autour de 4,5% dès 2030.[25] Cette rapide décroissance du quotient s'explique par l'arrivée de nombreux baby-boomers à l'âge de la retraite à partir des années 2000. Le pic de natalité d'après guerre expliqué par l'optimisme général, le redémarrage de l'économie et l'amélioration du niveau de vie global, a conduit à un accroissement démographique important allant de 1942 à 1973.

Or, un fort taux démographique entraîne une population active plus nombreuse, nécessitant ainsi d'importantes créations d'emplois pour garder constant le taux d'emploi au sein du pays. Seulement en temps de crise ou de ralentissement économique, les emplois se font rares et très peu sont créés, provoquant ainsi une hausse soudaine du nombre de chômeurs. Aussi, comme expliqué précédemment[26] le taux de chômage a tendance à baisser quand la population active diminue puisqu'il y a alors un moins grand écart entre l'offre et la demande d'emploi : la corrélation entre le taux de chômage et la population active est donc **positive**.

[24] https://data.oecd.org/fr/pop/taux-de-fecondite.htm (consulté le 09/02/2020)
[25] https://www.cairn.info/revue-de-l-ofce-2015-6-page-339.htm (consulté le 09/02/2020)
[26] Voir section I ; analyse économique du sujet page 8

GRAPHIQUE N°6 : Relation entre le taux de chômage et la population active



D'après le graphique n°6 on observe une relation positive comme supposée théoriquement, en revanche cette relation n'apparaît pas clairement pour la période 2016-2019 qui voit le taux de chômage diminuer (retour à une situation plus stable de l'économie française) quand la population active stagne autour de 29 700 milliers de personnes. Tandis qu'en moyenne chaque trimestre le flux d'entrées et de sorties du marché du travail est de 41 milliers de personnes (entrées > sorties), la population active a augmenté de 733 milliers de personnes entre le dernier trimestre de 2013 et le premier de 2014, ce qui est identifiable sur le graphique où l'on remarque un "trou" dans les observations de la période 2010-2015. Finalement on voit que les observations des périodes 2007-2009 et 2016-2019, périodes où le taux de chômage décroît (en tout cas entre 2007 et 2008 pour la couleur verte), ne sont pas alignées - de telle manière que l'on ne distingue pas la relation positive qui lie Y_t et X_4 sur ces périodes.

E) Évolution des effectifs dans l'industrie manufacturière : X_5

La Banque de France comme l'INSEE réalisent chaque mois des enquêtes de conjoncture prenant la forme d'indicateurs du climat des affaires, qui décrivent la situation conjoncturelle du mois passé, et prévoient à court terme le PIB trimestriel (grâce aux réponses de 10 000 dirigeants d'entreprises pour la Banque de France).[27] Ces enquêtes concernent de nombreux secteurs tels que l'industrie, les services, le bâtiment etc., et sont disponibles au niveau national comme régional. Étant donné la régularité et la rapidité de ces

[27] https://www.banque-france.fr/statistiques/conjoncture/enquetes-de-conjoncture (consulté le 16/05/2020)

enquêtes (le 15 du mois pour la Banque de France et entre le 27 et le 30 pour l'INSEE), elles constituent une source d'information importante enseignant les principaux diagnostics de l'économie française. De plus, ces données sont collectées auprès d'un échantillon de chefs d'entreprise - ce ne sont donc pas des indicateurs économiques à proprement parler, c'est-à-dire qu'ils ne sont pas simplement des taux (chômage, intérêts…) ni même des indices statistiquement calculés (IDH, Gini...). Ce sont plutôt des variables qualitatives, où sont agrégés les avis des chefs d'entreprises sur l'évolution de la conjoncture dans leur secteur, sous forme d'un indicateur synthétique d'opinion.[28] Depuis 50 ans, grâce à de telles enquêtes on retrouve notamment l'évolution du prix des matières premières, des produits finis, les variations des stocks, des dépenses d'investissement, de la production ou encore des effectifs. C'est ce dernier qui va nous intéresser dans le cadre de l'étude du taux de chômage.

Effectivement, l'évolution des effectifs représente l'évolution de la main d'œuvre dans les différents secteurs. Lorsqu'une entreprise augmente sa production elle a besoin de plus de main d'œuvre, elle embauchera donc de nouveaux employés pour satisfaire la demande globale en fortifiant ainsi ses effectifs. Une hausse des effectifs est donc synonyme d'amélioration et/ou de création d'emplois. Nous nous intéressons dans cette analyse aux évolutions des effectifs dans l'industrie manufacturière, récoltées au travers des enquêtes mensuelles de conjoncture (EMC) de la Banque de France. L'industrie est un secteur particulier car il a connu de nombreuses phases depuis la Révolution Industrielle émanant du Royaume-Uni dans la deuxième moitié du 18e, jusqu'à l'apparition de l'informatique et des systèmes de télécommunication dans les années 1980. Avec ces différentes phases, les secteurs ont beaucoup évolué - ils se regroupent aujourd'hui en 3 catégories[29] :

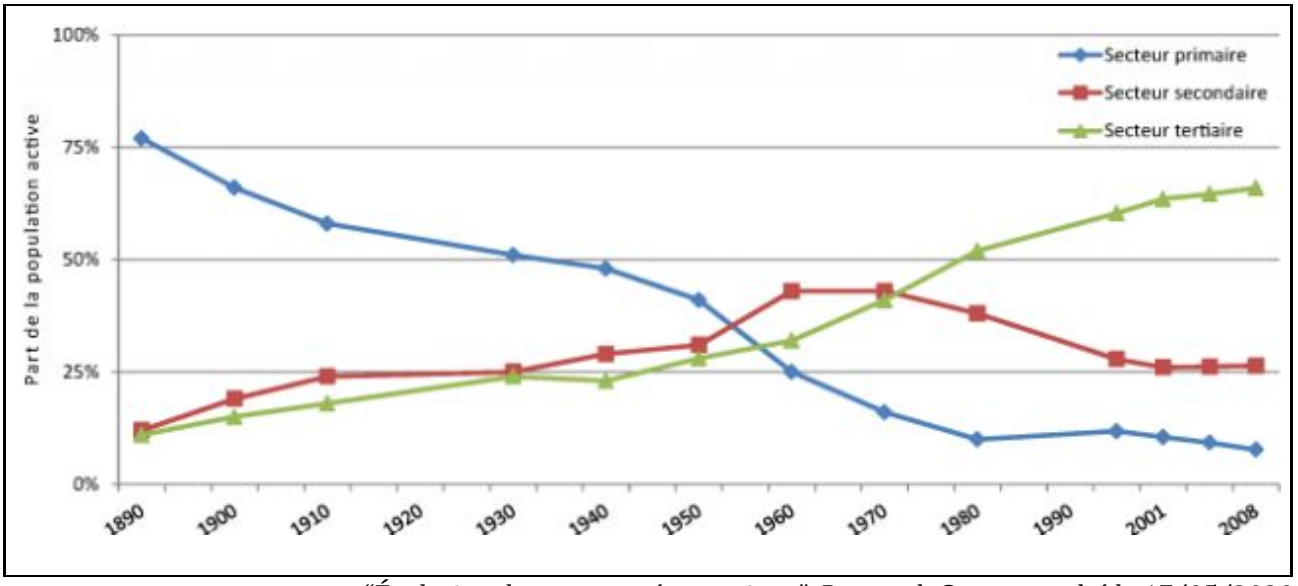
- le secteur primaire : regroupe les activités agricoles
- le secteur secondaire : regroupe les activités impliquant une transformation des matières premières (c'est-à-dire l'industrie)
- le secteur tertiaire : regroupe les services

Quoi qu'il en soit, depuis cette troisième "révolution" on assiste à une raréfaction du secteur secondaire où les emplois diminuent du fait de la substitution du capital au travail ; sur le graphique n°7 on constate un déclin du secteur industriel à partir des années 70. On

[28] https://www.senat.fr/rap/r02-254/r02-2549.html (consulté le 16/05/2020)
[29].https://www.vie-publique.fr/fiches/269995-les-grands-secteurs-de-production-primaire-secondaire-et-tertiaire (consulté le 17/05/2020)

voit aussi sur le graphique de l'annexe n°2 qui relate l'emploi industriel de 1980 à 2008, une diminution forte de l'emploi dans ce secteur, induisant une réduction importante de ses effectifs.

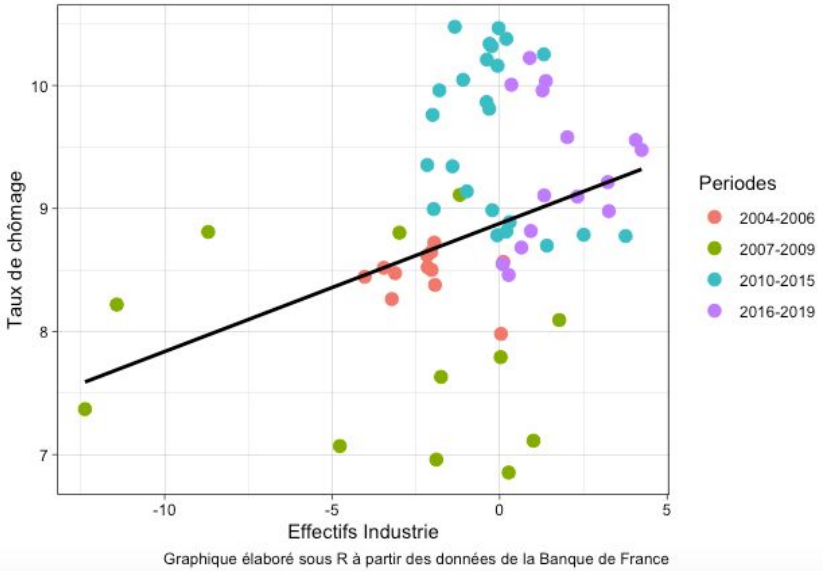
GRAPHIQUE n°7 : Évolutions des secteurs économiques



source : "Évolution des secteurs économique", Research Gate, consulté le 17/05/2020

Nous attendons une relation négative entre le taux de chômage et les effectifs dans l'industrie manufacturière, puisque lorsque les chefs d'entreprise estiment que les effectifs dans leur secteur ont augmenté par rapport au mois précédent, cela signifie que de nouveaux emplois ont été pourvus voire créés. Cela a pour conséquence d'améliorer l'emploi et donc de diminuer le taux de chômage. Regardons la nature de cette corrélation sur notre échantillon de données.

GRAPHIQUE n°8 : Relation entre le taux de chômage et les effectifs du secteur industriel



À partir du graphique n°8 on observe une relation apparemment positive entre le taux de chômage et les effectifs dans l'industrie, ce qui est contraire aux suppositions émises précédemment. Seulement, il apparaît que lorsque l'on prend chaque période individuellement, 2 sur 4 présentent en réalité une relation négative entre Y_t et X_5 comme nous pouvons le voir en annexe n°3. C'est donc encore une fois l'effet global qui l'emporte, et c'est la conjonction des sous périodes, et notamment les valeurs inhabituelles de la crise, qui donnent finalement une relation positive.

Partie 2 : Méthodologie économétrique

L'économétrie est la mesure de l'économie qui se fait grâce à la création et l'estimation de modèles ayant pour but d'identifier la relation entre des variables Y et X. Seulement, lors de telles modélisations, de nombreuses sources peuvent détériorer la qualité du modèle et l'estimation des paramètres. Il existe ainsi des méthodes d'estimation alternatives fondées sur les variations passées d'une série Y, de manière à minimiser les erreurs en se basant sur les éléments connus d'une série ; c'est l'étude des séries temporelles. Ainsi, une série temporelle ou 'processus temporel' est une suite d'observations numériques d'une variable donnée dans le temps, qui est caractérisée par une tendance (à la hausse ou à la baisse), des cycles, une saisonnalité et une composante accidentelle.

Dans cette deuxième partie nous allons expliquer la méthodologie que nous appliquerons en partie 3, dans le but de démontrer l'apport de l'outil Google Trends dans la prédiction de l'emploi et du chômage en France. Pour cela nous commencerons par modéliser le taux de chômage à travers un modèle ARIMA sans variables explicatives qui nous servira en quelque sorte de modèle 'benchmark'. Puis nous chercherons à cointégrer les variables explicatives avec Y_t pour construire différents modèles ARX constitués seulement des variables cointégrées, pour éviter toute relation fallacieuse qui puisse altérer les conclusions de notre analyse. Enfin, nous procéderons à la comparaison des modèles afin de voir si l'ajout de la variable issue de Google Tendances, améliore significativement les prévisions du taux de chômage. Nous réalisons les différentes analyses à l'aide du langage R, du logiciel Gretl ainsi que du programme JDemetra+.

I. Modélisations et prévisions ARIMA

C'est en 1927 que G.U.Yule introduit les modèles auto régressifs AR et que E.Slutsky introduit les modèles moyennes mobiles MA. En 1938, Wold propose un modèle ARMA combinant les modèles proposés par Yule et Slutsky, en un modèle linéaire basé sur la notion d'un processus infini de chocs aléatoires. Enfin, en 1954, il propose un processus ARMA stationnaire où :

➢ la partie auto régressive (AR) est constituée d'une combinaison linéaire finie de valeurs passées du processus et du terme aléatoire. Ce processus permet de modéliser les observations actuelles qui dépendent des observations antérieures.

➢ la partie moyenne mobile (MA) est constituée d'une combinaison linéaire finie de valeurs passées d'une variable aléatoire appelée "bruit blanc". Ce dernier est une suite de variables non corrélées à la variance constante et l'espérance mathématique nulle, il aide à représenter les effets d'un choc dans un futur proche.

Le processus ARMA a été popularisé en 1970 lorsque les chercheurs Box et Jenkins ont publié l'ouvrage "*Time series analysis, forecasting and control*", montrant que la méthodologie ARMA pour l'étude de séries temporelles pouvait s'appliquer à de nombreux domaines. Le modèle ARMA ainsi popularisé est composé de 'p' éléments pour déterminer AR et de 'q' éléments pour déterminer MA ; tels que ARMA[p,d].[30] On a donc ;

$$\text{AR(p)} : Z_t = \delta + a_t + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p}$$

$$\text{MA(q)} : Z_t = \mu + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

Aussi, il y a 2 conditions pour qu'un processus $[X_t]$ soit considéré comme un processus ARMA[p,d] :

➢ **<u>Stationnarité</u>** pour la partie AR, signifie que la série doit garder les mêmes caractéristiques dans le temps, on dit ainsi qu'elle est indépendante du temps. Ce concept de stationnarité est très important et constitue la première étape de la méthodologie ARIMA, il implique une même distribution des probabilités, une même espérance mathématique et une même variance pour chaque observation. Par conséquent, une série stationnaire tend à varier autour d'une valeur moyenne à long terme, c'est pourquoi la méthode d'estimation et de prévision 'Box Jenkins' s'applique à très court terme puisqu'à plus long terme c'est la moyenne mathématique qui dira les prédictions. Statistiquement, la condition se matérialise de la manière suivante ; pour un processus AR(1) on doit avoir $|\phi_1| < 1$ et pour un AR(2) on doit avoir $|\phi_2| < 1$, $\phi_1 + \phi_2 < 1$ et $\phi_2 - \phi_1 < 1$.[31]

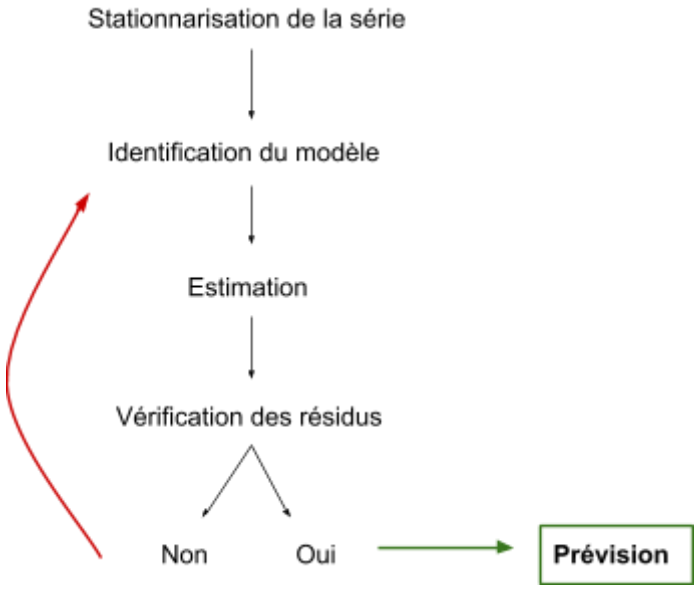
➢ **<u>Inversibilité</u>** pour la partie MA où tout est aléatoire, cette condition est indépendante de la stationnarité et est donc applicable sur un processus non stationnaire. De même que pour la stationnarité, la condition d'inversibilité se vérifie à travers la valeur des

[30] https://www.math.u-bordeaux.fr/~hzhang/m2/st/TS1.pdf (consulté le 13/02/2020)
[31] Tsay R. S., "Analysis of financial times series", *University of Chicago*, 2002, pp.28-38.

coefficients estimés par le processus MA ; il faut $|\theta_1| < 1$ pour un MA(1), et $|\theta_2| < 1$, $\theta_1 + \theta_2 < 1$ et $\theta_2 - \theta_1 < 1$ pour un MA(2) .[32]

Les chercheurs Box et Jenkins ont développé une méthode permettant de trouver les paramètres 'p' et 'q' d'un modèle ARMA[p,q][33] et tenter de prévoir les variations d'une série sur le court terme, dont le déroulé est visible en schéma n°2.

SCHÉMA N°2 : Méthode Box-Jenkins pour l'analyse de Y



source : élaboration propre à partir de l'outil "dessin" sous google drive

1- Stationnarisation de la série

Comme expliqué précédemment la première étape de la méthodologie ARIMA consiste à vérifier la stationnarité de la série, de manière à pouvoir identifier le modèle pour l'estimer et, s'il est vérifié, passer à la prévision. La stationnarité s'effectue à 2 niveaux : au niveau de la moyenne et au niveau de la variance.

➢ Stationnarité autour de la **moyenne** : la série doit fluctuer autour d'une valeur moyenne constante, *i.e.* indépendamment du temps.

➢ Stationnarité autour de la **variance** : la série doit avoir des variations de même ampleur autour de la moyenne, *i.e.* sa covariance ne doit pas dépendre du temps.

[32] Tsay R. S., "Analysis of financial times series", *University of Chicago*, 2002, pp.42-48.
[33] **Modèle ARMA** : Autoregressive moving average model

Un processus est dit stationnaire si sa variance et sa moyenne sont constantes dans le temps, cela signifie que la série ne comporte ni tendance, ni saisonnalité. Aussi, dans une série stationnaire l'autocovariance entre 2 observations dépendra uniquement du nombre de périodes qui les sépare. Il y a par conséquent 2 types de processus non stationnaires : les *Trend Stationnary* (TS) et les *Differency Stationnary* (DS).[34]

Pour vérifier la stationnarité d'une série on a recours à 2 types de test ; le test de Dickey-Fuller augmenté (ADF) ainsi que le test KPSS. Le test ADF est basé sur les hypothèses suivantes :

(1) H_0 : La racine unitaire existe, X_t n'est pas stationnaire

H_1 : X_t est stationnaire

(2) $t_{obs} \gtreqless ADF_{0,05}$

(3) Quand la p-value diminue on refuse H_0 ; X_t est stationnaire.

Le test KPSS quant à lui, repose sur les hypothèses inverses à celles du test ADF, c'est-à-dire que l'hypothèse nulle est H_0 : la série est stationnaire.

Prenons l'exemple d'une série brute Y_t que l'on cherche à modéliser à l'aide d'un processus ARMA. Imaginons que cette série ne soit pas stationnaire, on ne peut alors pas utiliser un modèle ARMA[p,q], il va falloir procéder à une transformation de la série brute Y_t. Si cette dernière n'est pas stationnaire autour de la variance, c'est-à-dire qu'il existe une tendance à la hausse ou à la baisse, on va alors la mettre en logarithme de manière à corriger cette *trend*, on a ainsi : y_t=log(Y_t). Si la série n'est pas stationnaire autour de la moyenne, c'est-à-dire que les dispersions autour de la moyenne fluctuent au cours du temps, on va alors procéder à une différenciation qui viendra corriger ces écarts. On sera alors dans le cadre d'un processus ARIMA[p,d,q] c'est-à-dire *Autoregressive integrated moving average model*, où I correspond au degré de différenciation de la série : I(0) si la série en niveau est déjà stationnaire.

2- Identification du modèle

La deuxième étape de la méthode de Box-Jenkins consiste à identifier le type de processus, cela est possible en regardant la forme des fonctions d'autocorrélation (FAC) et d'autocorrélation partielle (FACP) d'une série stationnaire (d'où l'importance de l'étape

[34] Bourbonnais R., "Économétrie", *DUNOD Editions*, 2018, pp.258-265.

précédente). Ces tracés sont contenus dans les corrélogrammes qui ont été développés pour la première fois en 1884 par le physicien Poynting, qui étudiait la relation entre le mouvement du prix du blé et les importations de coton et de soie.[35] Le processus d'identification du modèle est donc un processus visuel qui s'effectue uniquement en regardant la FAC et la FACP, 3 cas de figure sont alors possibles :

➢ Modèle classique : la ou les premières valeurs d'un des corrélogrammes sont significatives puis décroissent pour se placer autour de zéro à terme, et seules 1 ou 2 valeurs sont significatives dans l'autre corrélogramme, puis on observe une rupture. Cela signifie qu'il n'y a pas de partie saisonnière dans la série. Un modèle classique se présente sous forme d'un ARIMA[p,d,q] avec d=1 s'il a fallu différencier la série une fois pour la rendre stationnaire.

➢ Modèle saisonnier : la ou les premières valeurs des corrélogrammes ne sont pas significatives, en revanche sur les périodes postérieures on observe des valeurs significatives que ce soit dans la FAC ou dans la FACP. Ce phénomène de saisonnalité peut apparaître dans les séries temporelles qui ne sont pas annuelles, par exemple si l'on étudie les ventes de glaces mensuelles on observera un pic chaque année au moment de l'été : c'est la saisonnalité. Un modèle saisonnier se présente sous forme d'un SARIMA[P,D,Q $]_S$ où 'S' correspond à la périodicité de la série, ainsi pour des données mensuelles S=12, pour des données trimestrielles S=4 etc.[36]

➢ Modèle mixte : les premières valeurs des corrélogrammes suivent la description d'un modèle classique, mais au bout de k périodes on observe des valeurs qui sortent du seuil de significativité. Nous sommes alors dans le cas d'un modèle classique ET d'un modèle saisonnier. Aussi appelé 'modèle multiplicatif', le modèle mixte se présente ainsi : ARIMA[p,d,q]*[P,D,Q $]_S$.[37]

La spécification des processus doit suivre une règle importante : la règle de parcimonie. Celle-ci indique qu'il est mieux de choisir un modèle avec le moins de paramètres possibles, pour perdre le moins d'information et s'éloigner au minimum de la série initiale.

3- Estimation du modèle

L'estimation du modèle identifié à l'étape précédente se fait de manière informatique, en revanche il convient de prêter attention aux résultats de cette estimation. Un modèle bien

[35] https://www.math.u-bordeaux.fr/~hzhang/m2/st/TS1.pdf (consulté le 14/02/2020)
[36] **Modèle SARIMA** : Seasonal autoregressive integrated moving average model
[37] Tsay R. S., "Analysis of financial times series", *University of Chicago*, 2002, pp.48-53 .

spécifié est un modèle dont les paramètres sont pertinents, c'est-à-dire qu'ils doivent être significatifs, et justes ; il faut que leur valeur ne soit ni trop proche de 0, ni trop proche de 1 pour vérifier les conditions de stationnarité et d'inversibilité (on fixe donc la limite à 0,95).

4- Vérification des résidus

L'étape de vérification des résidus notés ($\widehat{a_t}$) est très importante puisqu'elle permet de valider ou non le modèle identifié en deuxième étape de la méthode Box-Jenkins, permettant ainsi d'améliorer le modèle si besoin. Pour être validé, un modèle doit avoir des résidus indépendamment distribués, cela implique que ces derniers suivent un processus 'bruit blanc' (au moins pour les 2 premières valeurs du corrélogramme, à savoir K=1 et K=2) et qu'ils suivent approximativement une loi normale. En effet en observant le corrélogramme des résidus, aucune valeur ne doit (théoriquement) dépasser le seuil de significativité fixé par le test de Bartlett à $\widehat{\varphi_K}(\widehat{a_t}) = \pm\, 1,96\sqrt{\frac{1}{T}}$. Il y a 2 manières de vérifier la blancheur des résidus :

➢ <u>En analysant chaque autocorrélation de manière indépendante</u> : la première valeur de la FAC($\widehat{a_t}$) doit être très éloignée du seuil de significativité car cela montre que la variance est faible et donc que le modèle est pertinent. Dans le cas contraire (premières valeurs de la FAC significatives) il faut retourner à l'étape d'identification du modèle, car si les résidus ne suivent pas un processus de 'bruit blanc' cela peut signifier que le modèle est incorrect, peu pertinent, que la saisonnalité n'a pas été prise en compte ou que la série n'a pas été stationnarisée.

➢ <u>En analysant les résidus dans leur ensemble</u> : grâce au test portmanteau, ou test de Ljung-Box qui s'applique de la manière suivante :

(1) H_0 : $\varphi_1 = \varphi_2 = \dots = \varphi_K = 0$ (hypothèse d'indépendance des résidus)

H_1 : dépendance des résidus

(2) $Q = T \sum_{k=1}^{K} \varphi_K^2(\widehat{a_t}) \sim \chi^2$

(3) Quand la valeur de Q augmente la p-value diminue et on refuse H_0.
Quand la valeur de Q diminue la p-value augmente et on accepte H_0.

Aussi, chaque inadéquation du modèle est signalée par l'accroissement de la valeur de Q, lorsque l'on compare plusieurs modèles on va donc choisir celui qui a la p-value la plus élevée, de manière à accepter H_0, pour que les résidus soient vérifiés.

De même, on procède à l'analyse de la distribution des résidus pour voir s'ils suivent approximativement une loi normale. Cette hypothèse est importante mais non déterminante dans la validation du modèle, c'est-à-dire que si des résidus ne suivent pas une loi normale en revanche ils sont bien indépendamment distribués, nous serons en mesure d'accepter et de valider le modèle dont ils sont issus.

Il existe un certain nombre de tests pour vérifier la normalité d'une distribution : le test de Kolmogorov-Smirnov, de Shapiro-Wilk, du χ^2, de Jacques-Bera, d'Agostino etc. Dans cette analyse nous utiliserons le test d'Agostino qui se base sur la transformation des skewness et kurtosis et apporte ainsi une évaluation complète, pas seulement sur l'asymétrie mais aussi sur l'aplatissement de la distribution. En outre, il présente de nombreux avantages par rapport aux tests traditionnellement utilisés pour la vérification de la normalité, il fonctionne de la manière suivante :

(1) H_0 : La variable suit une loi normale

H_1 : La variable ne suit pas une loi normale

(2) $K^2 = \gamma_1^2 + \gamma_2^2 \sim \chi^2$

(3) Quand la p-value est supérieure à $\alpha = 0,05$ on accepte H_0.

5- Prévisions

"All models are false, but some are useful" (G.Box, 1979). Cette phrase de Box montre que les processus de modélisation et d'estimation des paramètres d'une série sont toujours approximatifs, mais que certains modèles sont tout de même utiles notamment pour faire de la prévision.

Ainsi, après avoir modélisé le taux de chômage par un processus ARIMA, nous chercherons à prédire ses évolutions futures à 3 horizons différents : h=1, h=4 et h=8. En effet, nous utiliserons ces mêmes horizons pour les prévisions à l'aide de modèles ARX, dans le but de noter des différences de précision de Google Trends dans la prédiction de l'emploi à court, moyen et long terme. Nous utiliserons pour cela des boucles construites sous R pour prédire Y_t en incluant les prédictions dans les valeurs connues de la série à chaque itération, dans le but de minimiser les erreurs de prévision.

II. Cointégration des variables

Avant d'initier tout processus de cointégration des variables, nous vérifierons que nos séries ne comportent pas de partie saisonnière. En effet, si elle existe, la saisonnalité peut masquer certaines informations et notamment les tendances de long terme par la récurrence d'un phénomène tous les mois, trimestres etc. Nous la détecterons au travers d'une fonction "isSeasonal", contenue dans le package '*seastest*' sous R, qui regroupe les 5 tests connus pour détecter la saisonnalité. Son utilisation est simple ; la sortie de ce test répond à la question contenue dans le nom de sa fonction : *TRUE* ou *FALSE* si la série contient, ou ne contient pas de partie saisonnière.

La désaisonnalisation consiste à atténuer les tendances de court terme de manière à supprimer les effets déterministes pour faire ressortir les informations contenues dans une série. Il existe 3 types de méthodes de désaisonnalisation ; paramétriques, non paramétriques et semi-paramétriques. Nous utiliserons pour notre part les méthodes paramétrique et semi-paramétrique à travers TRAMO-SEATS et X13 ARIMA dont nous comparerons les résultats ; nous favoriserons le modèle ayant le moins de paramètres (*i.e.* règle de parcimonie), les coefficients d'erreur les plus faibles, mais aussi celui qui validera les conditions fondamentales des modèles ARIMA, à savoir normalité et blancheur des résidus. Lorsque nous aurons désaisonnalisé les variables qui le nécessitent, nous tenterons de les cointégrer au taux de chômage.

Le principe de cointégration apparaît dans les années 1980 grâce à l'économètre Clive William John Granger, puis la méthode se développe dans les années 90 avec la cointégration de plusieurs séries. Il s'agit de voir si celles-ci évoluent dans le même sens, ainsi, elles sont cointégrées si elles présentent les mêmes évolutions sur toute la période. Rappelons que chaque série temporelle est caractérisée par une composante déterministe et une composante aléatoire. Le but de la cointégration est de voir au bout de combien de différenciations la composante déterministe disparaît pour qu'il ne reste que la partie stochastique, afin de voir si celle-ci peut être représentée par un modèle ARIMA (ou ARMA si la série est déjà stationnaire). Par conséquent, on cherche à savoir si les erreurs peuvent être corrigées à court terme pour que les 2 séries trouvent une tendance commune à long terme : pour cela on regarde s'il y a une réduction ou un ajustement des erreurs par rapport à une tendance centrale.

Il y a 2 conditions à vérifier pour parler de séries cointégrées :

➢ Même ordre d'intégration des 2 séries : pour pouvoir cointégrer 2 séries il est impératif qu'elles soient intégrées du même ordre, c'est-à-dire qu'elles aient besoin du même nombre de différenciations pour être rendues stationnaires. Pour vérifier la **stationnarité** d'une série il faut valider 3 conditions ; l'espérance de la série est indépendante du temps, la constante est finie et indépendante du temps, et la covariance entre Y_t et Y_{t-k} est une fonction finie de K périodes.

➢ La relation n'est pas fallacieuse : il faut vérifier que la relation entre les 2 séries est pertinente, pour cela on utilise le test de causalité d'Engle-Granger. Une **relation fallacieuse** est caractérisée par un R^2 élevé et une valeur de la statistique 't' élevée. De plus, la régression peut être acceptée d'un point de vue statistique mais pas d'un point de vue économique car les résidus ne sont pas stationnaires, donc la relation ne peut pas être interprétée.

En outre, une relation n'est pas fallacieuse lorsque les variables sont cointégrées, pour tester la cointégration on doit être capable d'exprimer Y_t en fonction de X_t tel que $\widehat{Y}_t - \widehat{\alpha} - \widehat{\beta} X_t = \widehat{\varepsilon}_t$ où les résidus, notés $\widehat{\varepsilon}_t$, doivent être intégrés d'ordre 0 : $\varepsilon_t \sim I(0)$. En effet, l'écart entre les 2 séries doit être constant à travers le temps pour parler de cointégration.

Pour voir si l'on peut construire une combinaison linéaire entre 2 variables, c'est-à-dire si on peut les cointégrer, on procède par étapes selon la méthodologie d'Engle-Granger :

1) On commence par vérifier que les 2 séries étudiées soient intégrées du même ordre, pour cela on a recours aux tests ADF ou KPSS dont les règles de décision ont été développées dans la section précédente. Dans le cas d'une série Y_t en niveau non stationnaire, les variables explicatives doivent être intégrées du même ordre à savoir I(1) s'il a fallu la différencier une fois pour la rendre stationnaire autour de la moyenne.
2) Ensuite on effectue une régression linéaire simple pour vérifier la stationnarité des résidus de long terme, et sachant qu'à terme les séries évoluent autour d'une tendance centrale, on prend les variables brutes pour cette régression. On commence alors à vérifier la deuxième condition de la cointégration ; il s'agit de voir si la relation est fallacieuse ou non.
3) On cherche ainsi à vérifier la relation énoncée précédemment à savoir $\widehat{Y}_t - \widehat{\alpha} - \widehat{\beta} X_t = \widehat{\varepsilon}_t \sim I(0)$ c'est-à-dire voir si les résidus sont intégrés d'ordre 0. Pour

cela on applique une fois de plus les tests ADF ou KPSS, sur les résidus sauvegardés de la régression linéaire simple estimée à l'étape précédente. Pour rappel si les séries en niveau sont stationnaires, c'est-à-dire qu'elles ne nécessitent pas de différenciation, alors les résidus peuvent être intégrés du même ordre à savoir d=0.

4) Enfin, pour vérifier la pertinence de la relation entre Y_t et X_t on s'intéresse à la relation de court terme en construisant un <u>modèle à correction d'erreur</u> (MCE). Engle et Granger ont montré que toute série cointégrée peut être représentée par un MCE. Celui-ci sert à voir l'influence des observations les unes par rapport aux autres, c'est pour cela qu'il prend en compte les variables différenciées (car de toute façon à long terme les séries se stabilisent autour de la moyenne). Pour se faire, on utilise les résidus de la relation de long terme que l'on décale afin de voir s'il existe une correction entre les observations actuelles et celles qui précèdent. On obtient une relation de la forme suivante : $\Delta Y_t = \alpha_1 \Delta X + \delta(\widehat{Y_{t-1}} - Y_{t-1}) + \upsilon_t$ où δ correspond au coefficient de correction. Celui-ci doit être impérativement significatif et négatif pour qu'il y ait un retour de Y_t à sa valeur d'équilibre de long terme qui est déterminée par $\beta X_{t-1} + \alpha$. En effet, lorsque Y_{t-1} est supérieur à cette expression d'équilibre, il y aura une force de rappel vers l'équilibre de long terme uniquement si δ <0. Ainsi, le modèle à correction d'erreur permet de modéliser conjointement les dynamiques de court terme représentées par les variables différenciées une seule fois, et une dynamique de long terme représentée par les variables brutes. Enfin, le coefficient estimé de δ donne le pourcentage de déséquilibre entre les 2 séries qui sera corrigé chaque période, nous pouvons ainsi voir au bout de combien de temps les variables s'apparentent à 100%, c'est-à-dire que l'on retrouve une tendance commune.

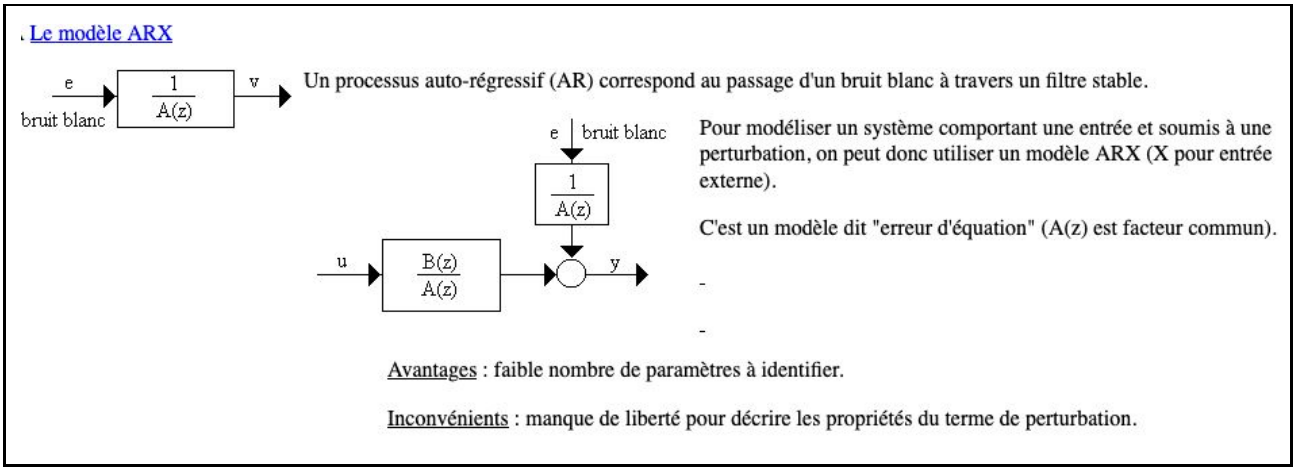
III. Modélisations et prévisions ARX

La cointégration des variables nous aura permis de voir lesquelles présentent des caractéristiques communes à la variable à expliquer et donc lesquelles sont intéressantes à utiliser dans des régressions avec variables explicatives. Les modèles ARX, c'est-à-dire *AutoRegressive with eXternal inputs models*, consistent en une modélisation de la composante aléatoire du processus, après extraction de la composante déterministe.[38] Ils sont donc

[38] https://tel.archives-ouvertes.fr/tel-00432051/document (consulté le 23/05/2020)

considérés comme des modèles de régression avec perturbations ARMA : il est ainsi possible d'utiliser l'information contenue dans les variables explicatives pour modéliser et prévoir le taux de chômage de manière plus précise.

SCHÉMA N°3 : Fonctionnement d'une modélisation ARX



source : http://francois.bateman.free.fr/HTML/filtrage-ident/chapitre%203/ident4/IDENT4.htm (consulté le 26/05/2020)

Comme nous le voyons sur le schéma n°3, le modèle ARX inclut des entrées 'u' et un bruit blanc 'e' de moyenne nulle. Nous réaliserons différents modèles prenant en compte toutes les variables explicatives cointégrées, chaque variable individuellement, ou encore en combinant certaines variables les unes aux autres. À partir de chacun de ces modèles nous réaliserons des prévisions qui nous permettront notamment de voir si la variable de Google Trends prise individuellement permet de prédire plus justement le taux de chômage que les autres variables exogènes prises individuellement aussi, ou au contraire, si l'outil est pertinent dans la prédiction de Y_t lorsqu'il est combiné aux indicateurs économiques traditionnels. Tout cela aux 3 horizons énoncés précédemment, dans le but de noter des changements sur le court, moyen ou long terme.

Étant donné que nous nous retrouverons avec un bon nombre de prévisions, nous les organiserons de la manière suivante : nous les noterons "prev*N*°H*N*°". En outre, les prévisions ARIMA constituent les premières prévisions réalisées et se noteront donc : **prev1H1** pour horizon à 1 période, **prev1H4** pour 4 périodes et **prev1H8** pour 8 périodes. Les prévisions du premier modèle ARX se noteront **prev2H1**, **prev2H4** etc.

IV. Comparaison des prévisions

L'objectif de cette étude étant de montrer l'apport de Google Tendances dans la prévision de l'emploi et du chômage, il est nécessaire de comparer les différentes prévisions construites à partir des modèles ARIMA comme ARX. Avant cela nous procéderons à une transformation des prévisions qui vise à les remettre au niveau de leur valeur initiale, dans le cas où nous avons différencié Y_t pour la rendre stationnaire. Puis nous procédons en 4 étapes :

➢ Représentations graphiques des prévisions : les graphiques permettent de visualiser un phénomène, ils permettront dans notre cas de se faire une première intuition pour savoir quel modèle paraît prédire le plus précisément le taux de chômage.

➢ Calcul des erreurs de prévision : à horizon h>0 il y a plusieurs moyens de connaître le pourcentage d'erreur associé à telle ou telle prévision. À partir de la différence entre les valeurs observées et les valeurs prédites, le pourcentage d'erreur peut être défini par les 3 critères suivants[39] :

- **L'erreur moyenne** : ce critère est très peu utilisé car la moyenne est faite sur les valeurs 'brutes' donc si l'erreur de prévision pour une année est de -0,3% et celle de l'année suivante est de +0,3% alors ce critère donnera une erreur moyenne de 0 c'est-à-dire une qualité de prévision parfaite, alors qu'en réalité elle ne l'est pas.
- **L'erreur absolue moyenne** : contrairement au premier critère, celui-ci fait la moyenne des erreurs de prévision en valeurs absolues, il corrige donc les limites de la méthode n°1.
- **L'erreur quadratique moyenne :** ce dernier est le critère le plus largement utilisé puisqu'il prend en compte la variance associée à chaque erreur de prévision en mettant chacune d'elles au carré et constitue donc le moyen le plus précis pour connaître l'erreur de prévision.

Ces deux dernières peuvent être calculées à partir de la médiane et non de la moyenne, elles offrent ainsi un complément pour connaître la qualité des prévisions. Ainsi, les meilleures

[39] Vaté M., "Statistiques chronologiques et prévisions", *Economica*, 1993, pp.217-225.

prévisions seront celles qui minimisent ces critères puisque cela reflète une estimation sans biais et de variance minimale.

➢ Application du test Diebold-Mariano : ce test permet de savoir si un modèle est statistiquement meilleur qu'un autre en évaluant sa capacité prédictive par rapport à un autre modèle. Développé en 1995 par Diebold et Mariano, le test DM s'applique de la manière suivante : si l'on cherche à comparer le modèle A au modèle B, alors on entre la ligne de commande suivante sous R : dm.test(residuals(modeleA),residuals(modeleB),alternative="l"). Le test suit la règle de décision suivante :

H_0 : Il n'y a pas de différence significative entre les deux modèles (autrement dit "les 2 modèles ont la même capacité prédictive")

H_1 : Avec l'alternative "less", le modèle B est moins précis que le modèle A

➢ Application du test multiple de Mariano et Preve : il permet de sélectionner le meilleur modèle en s'intéressant à la précision de tous les modèles, il sélectionne ainsi le ou les modèles ayant une capacité prédictive exceptionnelle en les classant le cas échéant. Il repose sur les hypothèses suivantes :

H_0 : Les prévisions évaluées ont la même précision

H_1 : L'Equal predictive accuracy (EPA) n'est pas atteinte

Toutes ces mesures de calcul, couplées aux graphiques de prévisions, nous permettront de conclure quant à l'apport de l'outil Google Trends pour prédire le chômage et l'emploi.

Partie 3 : Présentation des données et application

I. Présentation des données

TABLEAU N°2 : Échantillon de la base de données

	Taux de chômage	Popularité du mot 'emploi'	Taux d'intérêt	Production industrielle	Population active	Effectifs industrie
janv,-04	8,52	41,00	4,11	110,77	27 015,61	-3,46
avr,-04	8,38	39,33	4,31	110,97	27 033,46	-1,93
juil,-04	8,47	45,00	4,16	110,63	27 184,71	-3,12
oct,-04	8,50	38,00	3,83	111,26	27 178,72	-2,03
janv,-05	8,26	38,67	3,64	110,97	27 199,43	-3,22
avr,-05	8,44	36,00	3,37	110,88	27 324,99	-4,03
juil,-05	8,62	39,00	3,23	110,62	27 362,27	-2,16
oct,-05	8,65	32,33	3,39	111,63	27 326,01	-2,04
janv,-06	8,72	33,67	3,51	111,59	27 396,38	-1,95
avr,-06	8,57	29,67	3,99	113,42	27 396,62	0,11

Voici en tableau n°2 les 10 premières observations de la base de données élaborée dans le cadre de cette étude sur l'impact de Google Trends dans la prédiction du taux de chômage en France. Les observations sont trimestrielles et s'étendent de début 2004 au troisième trimestre de 2019, pour un total de 63 observations. Les données du taux de chômage, du taux d'intérêt, de la production industrielle et de la population active sont issues du site internet de l'OCDE.[40] Les données de popularité du mot 'emploi' quant à elles, sont issues du site Google Trends.[41] Enfin, comme stipulé dans la partie précédente, les données des effectifs dans le secteur industriel sont issues des EMC de la Banque de France. Le taux de chômage est exprimé en pourcentage de la population active, le taux d'intérêt en pourcentage par année, la production industrielle est indexée base 100 en 2015, la population active en milliers de personnes, la popularité du mot 'emploi' est calibrée entre 0 et 100 sur la période étudiée et les effectifs de l'industrie sont exprimés en solde d'opinion.

[40] https://data.oecd.org/fr/ (consulté le 26/01/2020)
[41] https://trends.google.fr/trends/?geo=FR (consulté le 23/01/2020)

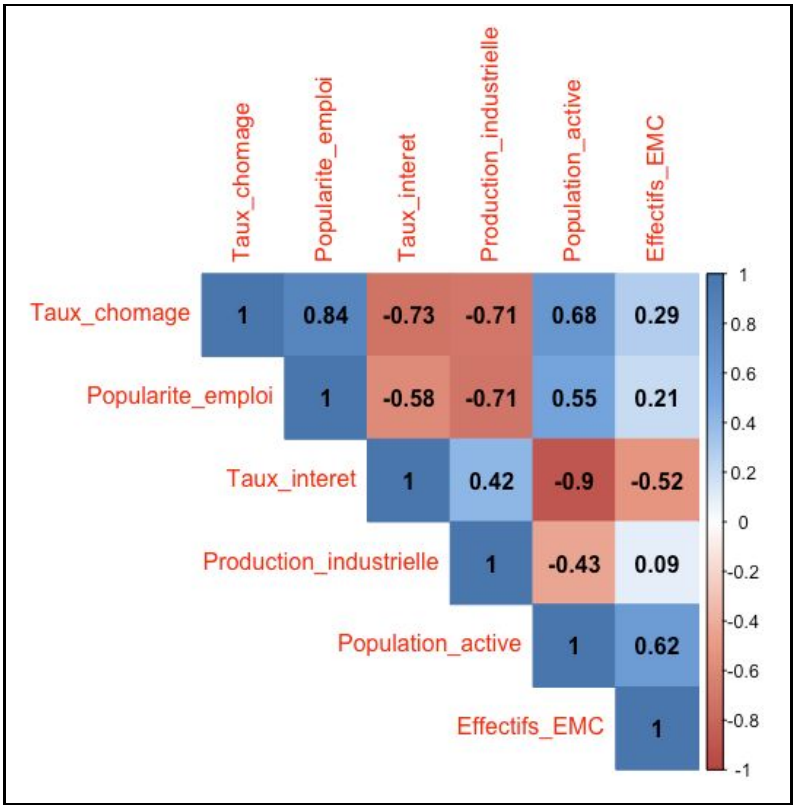
TABLEAU N°3 : Statistiques descriptives des 5 variables

```
> round(basicStats(df[,3:8]),2)
            Taux_chomage Popularite_emploi Taux_interet Production_industrielle Population_active Effectifs_EMC
nobs               63.00             63.00        63.00                   63.00             63.00         63.00
NAs                 0.00              0.00         0.00                    0.00              0.00          0.00
Minimum             6.85             24.67        -0.23                   93.57          27015.61        -12.38
Maximum            10.48             94.67         4.48                  115.38          29867.44          4.24
1. Quartile         8.51             38.84         0.89                   99.63          27800.50         -1.96
3. Quartile         9.79             82.50         3.71                  110.70          29503.47          0.92
Mean                8.98             61.39         2.50                  104.13          28531.30         -0.74
Median              8.82             65.33         3.01                  102.64          28324.70         -0.22
Variance            0.84            521.43         2.06                   35.57         823302.78          9.06
Stdev               0.92             22.83         1.44                    5.96            907.36          3.01
Skewness           -0.29             -0.26        -0.27                    0.43              0.03         -1.67
Kurtosis           -0.43             -1.41        -1.46                   -1.10             -1.43          4.44
```

Source : Logiciel R studio

D'après le tableau n°3 on voit que le taux de chômage varie entre 6,85% et 10,48% sur la période allant de 2004 à 2019, la médiane étant très proche de la moyenne cela montre qu'il ne semble pas y avoir de valeurs aberrantes ou atypiques qui tirent la moyenne à la hausse ou à la baisse, malgré ce que l'on aurait pu supposer avec la crise des Subprimes en 2008. L'écart type du taux de chômage est de 0,92 ce qui correspond à 10% de sa moyenne. La popularité du terme 'emploi' sous Google Trends fluctue entre 25 et 95%, elle n'atteint pas 100% parce que les données extraites de l'outil internet étaient initialement mensuelles, nous avons fait la moyenne tous les 3 mois pour les passer en trimestres. Cependant, les recherches atteignent leur maximum en septembre 2014 (c'est la région du Limousin qui rassemble le plus grand nombre de recherches parmi les régions de France). On note aussi que toutes les variables sont homogènes puisque les écart-types sont largement inférieurs aux moyennes sauf pour la variable des effectifs de l'industrie, cela se confirme par les diagrammes en moustache disponibles en annexe n°4 où l'on voit qu'aucune valeur des variables n'est atypique excepté X_5 pour laquelle il y en a 3. La moyenne de cette même variable est négative puisqu'elle se situe à -0,74 ; cela signifie que sur la période de 2004 à 2019 les effectifs dans l'industrie manufacturière ont diminué du fait de la désindustrialisation de nombreuses économies depuis la fin du 20e siècle, comme expliqué en partie précédente. Enfin, on voit à partir du skewness et de la kurtosis, que la variable dont la distribution s'apparente le plus à une loi normale est le taux de chômage. Celle qui s'en éloigne le plus est la variable des effectifs dans l'industrie - les histogrammes de distribution des variables, quant à eux, se trouvent en annexe n°5. De même, les résultats du test de Shapiro (basé sur l'hypothèse nulle : la variable suit une loi normale) appliqué à l'ensemble de notre base, révèlent que seule la variable du taux de chômage suit une loi normale au seuil de 1%.

FIGURE N°1 : Matrice de corrélation



Nous nous intéressons à présent à la matrice de corrélation. La corrélation entre les variables explicatives et le taux de chômage est visible en première ligne de la matrice ; on constate que la corrélation la plus forte est celle qui lie la popularité du mot "emploi" au taux de chômage, ce qui est bon signe pour notre analyse sur son apport prévisionnel. Les variables explicatives sont incidemment classées de la plus grande, à la plus petite corrélation pour laquelle le coefficient s'élève à 0,29 et correspond aux effectifs de l'industrie manufacturière.

TABLEAU N°4 : Valeurs atypiques des séries détectées sous JDemetra+

Taux de chômage : Y_t

	Period	Value	StdErr	TStat
LS	I-2009	0,7945	0,1998	3,9764

Popularité du mot 'emploi' : X_1

	Period	Value	StdErr	TStat
LS	I-2016	-10,6176	2,8535	-3,7209

Production industrielle : X_3

	Period	Value	StdErr	TStat
LS	IV-2008	-8,5014	1,0852	-7,8340
TC	I-2009	-8,6071	1,0388	-8,2855
TC	I-2011	3,7062	1,0214	3,6287

Population active : X_4

	Period	Value	StdErr	TStat
LS	I-2014	596,3298	81,3345	7,3318

La détection des points atypiques est importante dans l'étude de séries temporelles parce qu'elle permet de se rendre compte de l'impact de chocs exogènes sur les variables, que ce soient des indicateurs économiques ou autres. Le logiciel JDemetra+ a détecté des valeurs atypiques pour 4 de nos 6 variables : le taux de chômage, la variable Google Trends, la production industrielle et la population active. Sur les 6 points atypiques décelés, 4 sont de type *Level Shift* c'est-à-dire que l'effet du choc sur la série est permanent tandis que 2 sont des *Temporary Change* ce qui signifie que la série revient à son niveau d'origine.

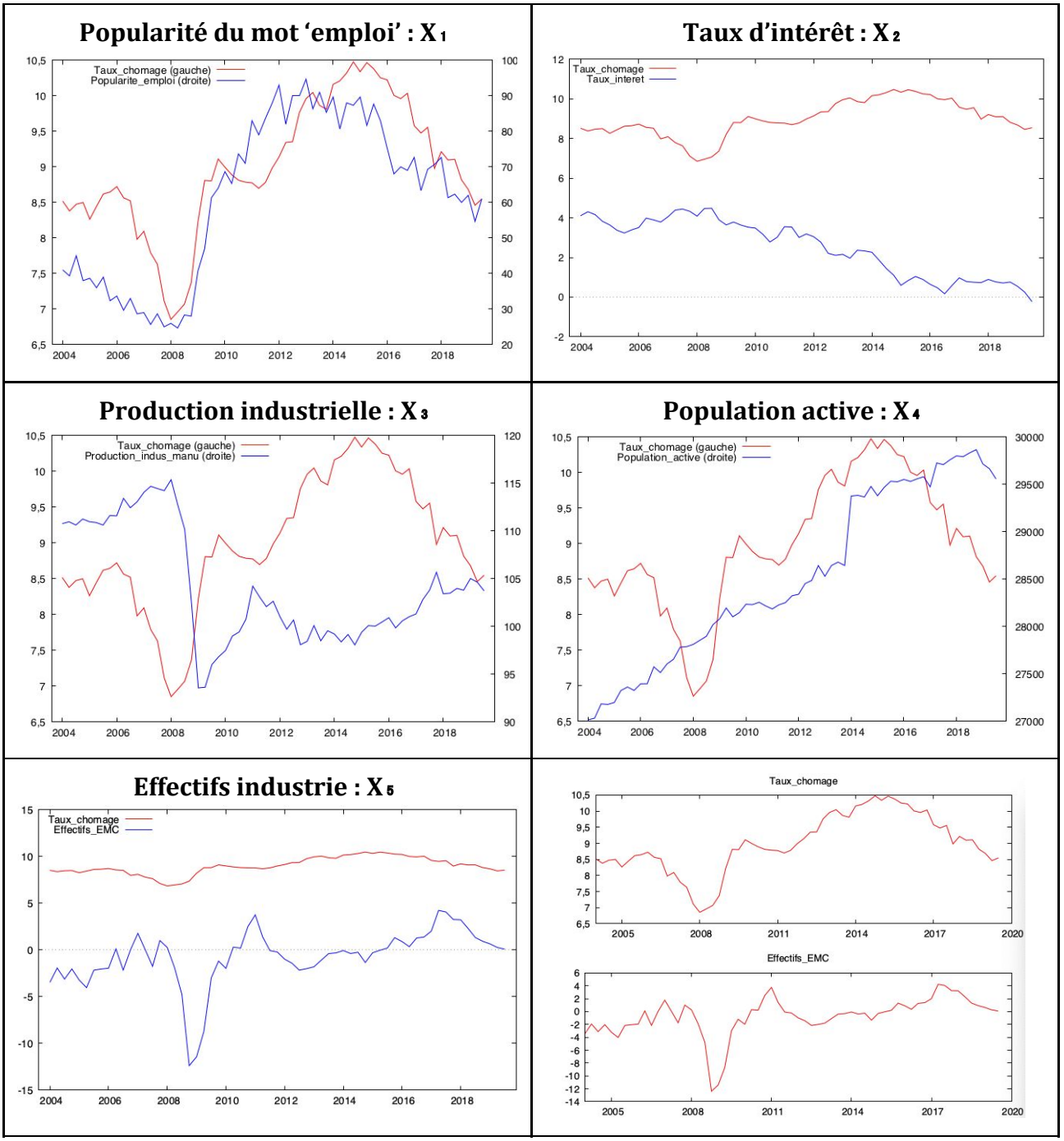
Le point atypique du premier trimestre de 2009 pour le chômage est imputable à la crise des Subprimes, pour la variable Google Trends il s'agit d'une baisse soudaine des recherches du terme 'emploi' qui est considérée comme atypique en janvier 2016. Effectivement, Pôle Emploi a recensé 27 900 demandeurs d'emploi de moins en janvier 2016 par rapport à décembre 2015, pour une baisse de 0,8%.[42] Il est intéressant de noter, encore une fois, comment les fluctuations de recherche sous Google Trends reflètent le marché du travail, et notamment l'offre de travail des demandeurs d'emploi.

Entre le premier trimestre de 2008 et le premier de 2009, la production industrielle (basée à 100 en 2015) passe de 115,38 à 93,57, soit une chute de près de 22 points en seulement 1 an. Celle-ci redémarre à partir du deuxième trimestre de 2009 jusqu'au premier de 2011 où elle atteint un pic puis s'affaiblit jusqu'en 2014. Ainsi, les 3 points atypiques de la production industrielle sont-ils liés à la chute et la reprise soudaine de l'activité. Enfin, le point considéré comme atypique pour la variable "population active" correspond à la hausse subite du nombre de personnes sur le marché du travail ; le flux d'entrées et de sorties était plus de 17 fois plus important que le flux moyen sur la période 2004-2019.

Nous pouvons terminer la présentation et l'étude des variables en regardant l'évolution dans le temps de chacune d'entre elles mise sur un même graphique que celui du taux de chômage (Y_t), de manière à voir si elles ont des évolutions similaires et ainsi pressentir si elles pourront être cointégrées ou non.

[42].https://travail-emploi.gouv.fr/archives/archives-presse/archives-communiques-de-presse/article/les-demandeurs-d-emploi-en-janvier-2016 (consulté le 22/05/2020)

FIGURE N°2 : Graphiques d'évolution des variables explicatives



Sur la figure n°2 on voit en rouge le taux de chômage et en bleu les 5 variables explicatives. On constate d'emblée que les variables X_1 et X_3 suivent approximativement les fluctuations du taux de chômage, on peut donc facilement supposer qu'elles seront cointégrées à ce dernier puisqu'elles convergent à long terme. En revanche, les variables X_2 et X_4 semblent évoluer d'une manière totalement indépendante au taux de chômage, elles n'ont pas les mêmes variations sur la période étudiée, il apparaît ainsi qu'elles seront difficilement co-intégrables. De plus on note que les fluctuations du taux d'intérêt sont exactement

opposées à celles du taux de chômage, il y a comme une symétrie de leurs évolutions ce qui confirme la relation négative que l'on a pu observer entre ces 2 variables sur le graphique de corrélation n°3. Concernant les effectifs dans l'industrie, si l'on regarde le graphique qui met sur une même échelle les 2 variables, il apparaît qu'elles ne suivent pas les mêmes évolutions dans le temps. Par ailleurs, si l'on s'intéresse à leurs fluctuations avec chacune une échelle différente, on s'aperçoit que les variations sont similaires ; on peut alors penser que cette variable sera cointégrée au taux de chômage. Nous pourrons vérifier ces hypothèses par la suite, dans la troisième section de cette partie.

Le tableau qui suit reprend les principaux éléments qui caractérisent les variables.

TABLEAU N°5 : Principales caractéristiques des variables de cette étude

	Taux de chômage	**Popularité du mot 'emploi'**	**Taux d'intérêt**	**Production industrielle**	**Population active**	**Effectifs industrie**
Nom	Y_t	X_1	X_2	X_3	X_4	X_5
Source	OCDE	Google Trends	OCDE	OCDE	OCDE	Banque de France
Unité	% de la population active	compris entre 0 et 100	% par an	2015=100	milliers de personnes	solde d'opinion
Corrélation théorique*	/	positive	positive	négative	positive	négative
Corrélation empirique*	/	positive	négative	négative	positive	positive
Distribution normale	oui	non	non	non	non	non
Valeurs atypiques	1	1	0	3	1	0
Cointégration possible	/	✓	✗	✓	✗	✓

* : corrélations entre les variables explicatives et Y_t

II. Modélisations et prévisions ARIMA

Nous allons dans cette section suivre la méthodologie de Box et Jenkins développée dans la partie précédente dans le but de modéliser le taux de chômage sans variables explicatives, il nous servira donc de modèle de référence dit “benchmark” pour voir si l’ajout de variables exogènes améliore la qualité des prévisions. Nous effectuerons les recherches de modèle avec le logiciel Gretl, puis nous ferons les prévisions à l’aide du langage R.

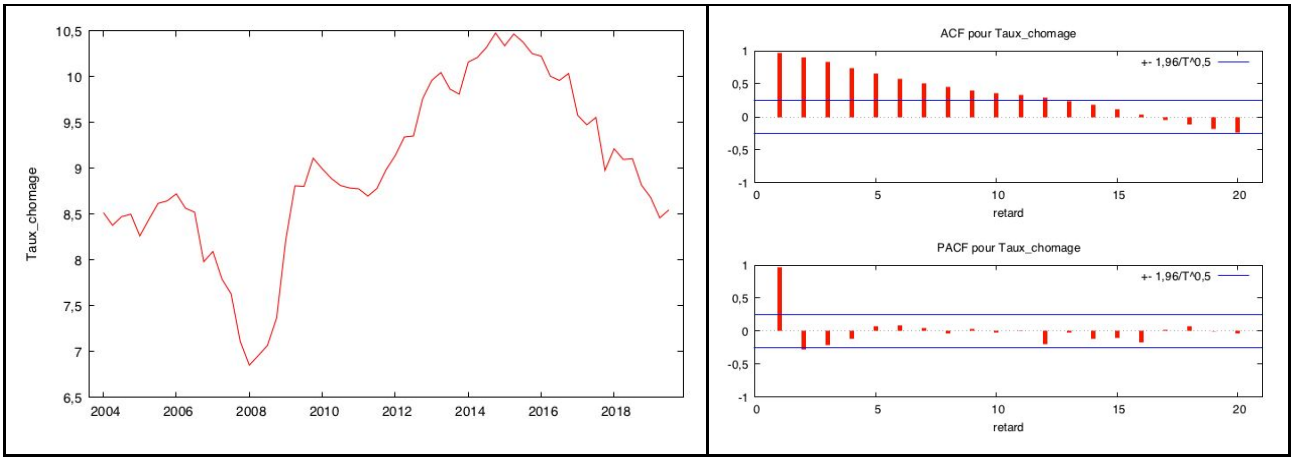
Notre série étant trimestrielle nous pouvons supposer l’existence d’une saisonnalité où l’on retrouve un même phénomène toutes les 4 périodes. De plus, les observations s’étendant de 2004 à 2019, nous pouvons imaginer que la crise mondiale de 2008 créera de fortes fluctuations dans notre série puisqu’il s’agit d’un indicateur économique.

Nous décidons de prendre un ordre maximum de retard de 20 périodes pour les corrélogrammes, ce qui correspond à 5 années d’observations consécutives allant de début 2004 à fin 2008. En prenant un nombre important de retards nous retenons l’impact de la crise de manière à garder un échantillon représentatif, et à visualiser d’éventuelles valeurs significatives en dehors des premières périodes dans les différents corrélogrammes.

A) Stationnarisation de la série

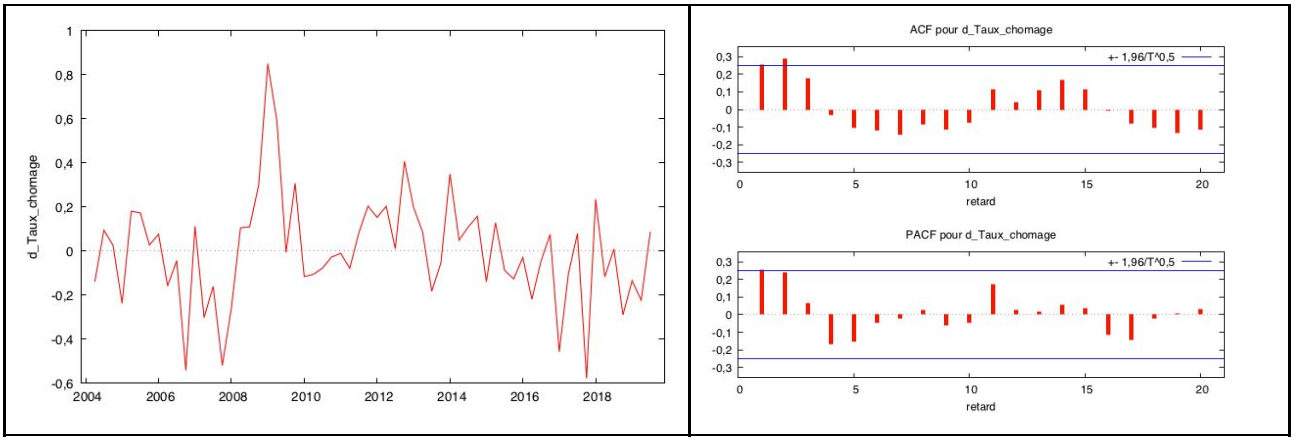
Comme expliqué précédemment, un processus est dit stationnaire si la série est stable pour la variance et pour la moyenne ; c'est-à-dire qu’il ne doit pas y avoir de tendance à la hausse ou à la baisse, ni de coupure dans la série ou encore de processus aléatoire.

GRAPHIQUE N°9 : Évolution temporelle et corrélogramme du taux de chômage brut (Y_t)



Il apparaît sur le graphique n°9 que la série n'est pas stationnaire, en effet les dispersions autour de la moyenne augmentent et diminuent fortement sur la période étudiée avec un pic inhabituel lié à la crise de 2008, comme supposé précédemment. Il convient donc de passer par un retard pour corriger cette trop grande volatilité de la série, en faisant une différenciation qui nous coûtera une observation (nous n'aurons plus que 62 périodes), mais permettra de stabiliser la série. Il ne semble pas y avoir de tendance à la hausse ou à la baisse car la chute du taux de chômage jusqu'à 2008 est compensée par une forte hausse dès 2011, il n'est donc pas nécessaire de passer la série en logarithme car elle est déjà stationnaire autour de la variance. Nous effectuons donc une différenciation sur la série initiale telle que $Z_t = Y_t - Y_{t-1}$. Nous passons donc d'une série composée de 63 observations à une série de 62 observations, qui commence au deuxième trimestre de 2004 au lieu du premier. On note aussi que les valeurs dans la fonction d'autocorrélation décroissent lentement tandis qu'on observe une réelle coupure dans la fonction d'autocorrélation partielle dès K=2 ; on peut supposer que le taux de chômage est modélisable par un processus auto régressif AR(p). Enfin, nous constatons qu'il n'y a pas de valeurs significatives en dehors des premières valeurs, cela signifie qu'il n'y a pas de saisonnalité, contrairement à ce que l'on avait supposé avant de regarder les graphiques d'évolution.

GRAPHIQUE N°10 : Évolution temporelle et corrélogramme du taux de chômage différencié



Après une différenciation visible sur le graphique n°10, il semble que la série soit davantage stationnaire par rapport à la moyenne ; l'inertie générale de la série a été réduite, les dispersions autour de la moyenne sont plus homogènes. La transformation était donc nécessaire pour stationnariser la série, l'effet de la crise des Subprimes est toujours visible mais semble atténué puisque l'inertie post-choc est plus faible : la trend revient rapidement à

la moyenne. Notre série Y_t est stationnarisée grâce à une différenciation, nous la notons Z_t et l'utiliserons pour la méthode Box-Jenkins. Enfin, nous pouvons vérifier la stationnarité de Z_t en effectuant le test ADF dont la règle de décision a été énoncée dans la partie de méthodologie économétrique. La p-value associée au test de la racine unitaire pour notre modèle est 0,0003 comme visible dans le tableau n°6, nous refusons H_0 ; notre série différenciée une fois est bel et bien stationnaire. Nous pouvons à présent passer à l'estimation en cherchant le modèle correspondant à cette série.

TABLEAU N°6 : Résultats du test ADF sous Gretl

```
Test de Dickey-Fuller augmenté pour d_Taux_chomage
testing down from 10 lags, criterion AIC
taille de l'échantillon 60
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec un retard de (1-L)d_Taux_chomage
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,559901
  statistique de test: tau_nc(1) = -3,60912
  p. critique asymptotique 0,000304
  Coeff. d'autocorrélation du 1er ordre pour e: -0,018
```

B) Identification et estimation du modèle

L'annexe n°6 indique que les 2 premières valeurs de la FAC et la première de la FACP de Z_t sont significatives au seuil de 5%. Sachant qu'aucune autre valeur n'est significative au bout de K périodes, on sait qu'il s'agit d'un modèle classique qui se présentera donc sous la forme d'un ARIMA[p,d,q] où d=1 puisque nous avons différencié une fois la série initiale pour la rendre stationnaire. On observe à partir du corrélogramme du graphique n°10 qu'il y a 2 valeurs significatives dans la FAC et une dans la FACP ce qui nous amène à modéliser un processus ARIMA[1,1,0] puisque les corrélogrammes s'apparentent à ceux d'une spécification AR(1).

TABLEAU N°7 : Estimation d'un premier modèle ARIMA[1,1,0]

```
Évaluations de la fonction : 14
Évaluations du gradient : 6

Modèle 1: ARMA, utilisant les observations 2004:2-2019:3 (T = 62)
Estimated using AS 197 (MV exacte)
Variable dépendante: d_Taux_chomage
Écarts type basés sur la matrice hessienne

             coefficient   erreur std.      z       p. critique
  ---------------------------------------------------------------
  const      0,000202342   0,0397780    0,005087    0,9959
  phi_1      0,251465      0,122154     2,059       0,0395    **

  Moy. var. dép.             0,000482
  Éc. type var. dép.         0,245735
  Moyenne des innovations    0,000637
  Ec. type des innovations   0,235715
  R2                         0,064824
```

```
R2 ajusté                  0,064824
Log de vraisemblance       1,591349
Critère d'Akaike           2,817303
Critère de Schwarz         9,198706
Hannan-Quinn               5,322803

                      Réel  Imaginaire   Modulo   Fréquence
-----------------------------------------------------------
AR
Racine  1           3,9767     0,0000     3,9767     0,0000
-----------------------------------------------------------
```

L'estimation du modèle est disponible dans le tableau n°7, on voit d'emblée que la constante n'est pas significative, ce qui signifie que la série évolue autour de 0, il faut donc réestimer le modèle en retirant cette dernière.

TABLEAU N°8 : Estimation du même modèle ARIMA[1,1,0] sans la constante

```
Évaluations de la fonction : 16
Évaluations du gradient : 3

Modèle 2: ARMA, utilisant les observations 2004:2-2019:3 (T = 62)
Estimated using AS 197 (MV exacte)
Variable dépendante: d_Taux_chomage
Écarts type basés sur la matrice hessienne

            coefficient   erreur std.     z     p. critique
  ---------------------------------------------------------
  phi_1       0,251468      0,122152     2,059     0,0395    **

  Moy. var. dép.             0,000482
  Éc. type var. dép.         0,245735
  Moyenne des innovations    0,000789
  Ec. type des innovations   0,235715
  R2                         0,064827
  R2 ajusté                  0,064827
  Log de vraisemblance       1,591336
  Critère d'Akaike           0,817329
  Critère de Schwarz         5,071598
  Hannan-Quinn               2,487662

                      Réel  Imaginaire   Modulo   Fréquence
  -----------------------------------------------------------
  AR
  Racine  1           3,9766     0,0000     3,9766     0,0000
  -----------------------------------------------------------
```

On voit dans le tableau n°8 l'estimation de ce même modèle sans la constante, on peut voir que $\widehat{\phi_1}$ =0,251 donc les conditions de stationnarité et d'inversibilité sont respectées puisque $\widehat{\phi_1}$ est différent de 0, inférieur à 1 et en est assez éloigné pour être certain de la stationnarité. S'agissant d'un processus AR(1) auquel on ajoute une différentiation, celui-ci est par définition inversible. Économiquement parlant, $\widehat{\phi_1} = 0,251$ signifie que la série est stationnaire avec une légère persistance de 25% (sachant que 0< ϕ <1 où 0 signifie aucune persistance ou un retour immédiat à une situation normale, et où 100% signifie que la série ne reviendra jamais à son état initial). De plus, l'écart-type s'élève à 0,236 ce qui est peu comparé à l'écart-type de la variable brute de 0,92, et donc satisfaisant pour notre modèle.

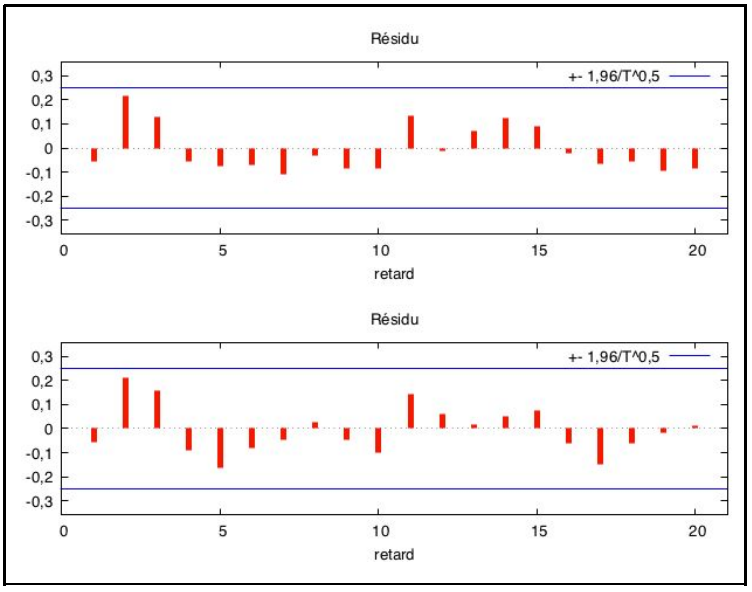
Nous pouvons passer à présent à l'étape de vérification des résidus qui viendra valider ou invalider ce modèle.

C) Vérification des résidus

Comme précisé en partie précédente, il existe 2 approches pour vérifier la blancheur des résidus. Nous commencerons par l'approche individuelle puis nous étudierons les résidus dans leur ensemble en regardant la valeur de la statistique Q. Enfin, nous regarderons la normalité des résidus grâce aux histogrammes et aux graphiques QQ, tout en privilégiant la valeur du Q-stat.

➢ Analyse individuelle : Sur le graphique n°11 on voit d'emblée que la première valeur du corrélogramme des résidus est très éloignée du seuil de significativité : elle se situe à -0,05 environ alors que le seuil, lui, est de 1,96 $\sqrt{\frac{1}{62}}$ =0,249. Cela indique que la variance est faible, donc notre modèle est pertinent. De plus, aucune valeur ne dépasse le seuil de significativité [-0,249 ; +0,249] donc les résidus suivent bien un processus 'bruit blanc' : ils sont stationnaires et stochastiques.

GRAPHIQUE N°11 : Corrélogramme et test des résidus du modèle ARIMA[1,1,0]



➢ Analyse globale : D'après le tableau n°9 on constate que la statistique Q du 20e retard s'élève à 13,39 pour une p-value de 0,818> $\alpha = 0,05$ ce qui signifie que l'on accepte largement l'hypothèse nulle selon laquelle les résidus sont indépendamment distribués. De même, le test de Box-Ljung appliqué à l'ensemble des résidus confirme leur indépendance par sa p-value égale à 0,64. Cette nouvelle approche nous permet encore une fois de valider le modèle ARIMA[1,1,0] puisque ses résidus ont été vérifiés. Enfin, nous pouvons vérifier la normalité de ces derniers.

TABLEAU N°9 : Analyse de la blancheur des résidus du modèle ARIMA[1,1,0]

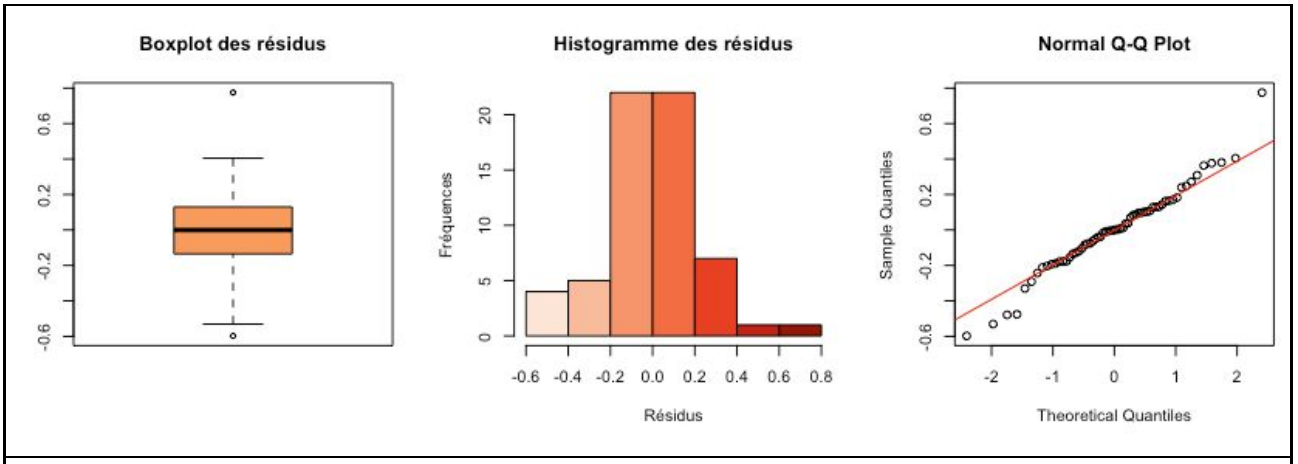
Fonction d'auto-corrélation résiduelle
***, **, * indicate significance at the 1%, 5%, 10% levels
using standard error 1/T^0,5

RETARD	ACF	PACF	Q	[p. crit.]
1	-0,0578	-0,0578		
2	0,2145 *	0,2118 *	3,2593	[0,071]
3	0,1304	0,1601	4,4025	[0,111]
4	-0,0567	-0,0906	4,6225	[0,202]
5	-0,0767	-0,1626	5,0325	[0,284]
6	-0,0684	-0,0807	5,3643	[0,373]
7	-0,1081	-0,0461	6,2068	[0,400]
8	-0,0316	0,0255	6,2803	[0,507]
9	-0,0867	-0,0454	6,8437	[0,554]
10	-0,0850	-0,1003	7,3945	[0,596]
11	0,1342	0,1422	8,7954	[0,552]
12	-0,0139	0,0587	8,8106	[0,639]
13	0,0703	0,0154	9,2108	[0,685]
14	0,1257	0,0514	10,5159	[0,651]
15	0,0886	0,0767	11,1791	[0,672]
16	-0,0201	-0,0604	11,2140	[0,737]
17	-0,0643	-0,1510	11,5782	[0,772]
18	-0,0579	-0,0631	11,8808	[0,807]
19	-0,0944	-0,0160	12,7035	[0,809]
20	-0,0849	0,0106	13,3850	[0,818]

```
> Box.test(residuals(ARIMA110), type = c("Ljung-Box"))

        Box-Ljung test

data:  residuals(ARIMA110)
X-squared = 0.21741, df = 1, p-value = 0.641
```

➢ Normalité des résidus :

GRAPHIQUE N°12 : Analyse de la distribution des résidus du modèle ARIMA[1,1,0]




```
> Agostino
                        Stat.     p-value
Omnibus D'Agos.   3.9474675 0.13893713
Skewness D'Agos.  0.3411145 0.73301741
Kurtosis D'Agos.  1.9573218 0.05030965
```

Le diagramme de Tukey, l'histogramme et la droite d'Henry nous montrent que les résidus du modèle ARIMA[1,1,0] comportent 2 valeurs atypiques ; un résidu a une valeur largement supérieure au 3è quantile et l'autre inférieur au premier. Ces points se retrouvent sur l'histogramme où l'on constate que la distribution est asymétrique à gauche, donc étalée à droite car le résidu atypique supérieur tire la distribution vers la gauche. De même, nous retrouvons ces valeurs sur le QQ plot puisqu'elles ne sont pas alignées sur la droite d'Henry. Le test d'Agostino nous informe que le coefficient d'asymétrie (skewness) est positif et significatif ; il y a donc une asymétrie à gauche comme nous l'avions remarqué. Cependant, le coefficient d'aplatissement n'est pas significatif à 5%, donc l'aplatissement plus faible de notre distribution par rapport à la loi normale n'est pas révélateur. Enfin, la ligne "Omnibus D'Agos." donne la valeur de la probabilité associée au test d'Agostino, sachant qu'elle est supérieure à 5% nous acceptons l'hypothèse nulle selon laquelle les résidus suivent une loi normale.

Ainsi, le modèle ARIMA[1,1,0] est un modèle pertinent que l'on valide. L'information liée à l'estimation et à la vérification de ce modèle se trouve dans le tableau n°10.

TABLEAU N°10 : Principales caractéristiques du modèle estimé

	ARIMA(1,1,0)
Nombre de coefficients significatifs*	1/1
Valeur du coefficient estimé	$\widehat{\phi}_1$ =0,251
Ecart-type	0,236
Log de vraisemblance	1,59
P-value blancheur des résidus	0,641
P-value normalité des résidus	0,139
Conclusion	✓

* : au seuil de risque de 5% tel que α =0,05

D) Prévisions

Nous réalisons dans cette section les prévisions du taux de chômage aux 3 horizons énoncés précédemment à savoir h=1, h=4 et h=8 - c'est-à-dire pour un trimestre, un an et

deux ans. Comme nous l'avons souligné précédemment, la série brute du taux de chômage n'est pas stationnaire, nous utiliserons donc la série différenciée pour les toutes les prévisions de cette étude. Avant de comparer les différentes prévisions nous procéderons à leur transformation pour les remettre au niveau de la série initiale, de manière à ne pas fausser les erreurs de prévisions par exemple.

➢ **Un point sur la série différenciée** : la variable du taux de chômage différencié fluctue entre -0,577 et 0,85 avec une moyenne de 0,0005 et une médiane de 0,0015. La moyenne est tirée vers le bas du fait des valeurs atypiques de la série qui sont toujours identifiables sur le graphique de la série stationnaire (confère graphique n°10 page 43).

TABLEAU N°11 : Prévisions à partir du modèle linéaire ARIMA[1,1,0]

	h=1	**h=4**	**h=8**
T4 2017			0.03954410
T1 2018			0.02560628
T2 2018			0.02081030
T3 2018			0.01915938
T4 2018		0.009285309	0.01859061
T1 2019		0.009300995	0.01839437
T2 2019		0.009304651	0.01832643
T3 2019	-0.04034225	0.009305291	0.01830269

Ainsi, à partir du modèle ARIMA[1,1,0] la différence du taux de chômage entre le deuxième et le troisième trimestre de 2019 serait de -0,0403. En réalité cette dernière se situait à 0,0879, il y a donc un écart de 0,1282 point. En général le modèle ne prévoit pas beaucoup de fluctuations dans les valeurs futures de la série du taux de chômage différencié, cela est compréhensible lorsque l'on regarde le graphique des valeurs observées et prédites par le modèle, disponibles en annexe n°7. Ce dernier modélise mal les variations de la série ; ainsi les valeurs prédites fluctuent entre -0,136 et 0,214 soit 4 fois moins d'amplitude que les valeurs observées.

Nous calculerons les erreurs de prévision de chacune des prévisions et de chacun des modèles lorsque nous comparerons toutes les prévisions en section V de cette partie.

III. Cointégration des variables

Nous allons dans cette section appliquer la méthode de cointégration dont la méthodologie a été développée antérieurement. Nous chercherons à cointégrer nos 5 variables en commençant par vérifier l'existence d'une saisonnalité, auquel cas nous l'enlèverons de manière à obtenir des séries corrigées des variations saisonnières (CVS).

A) Test de saisonnalité

Le test contenu dans le package *'seastests'* nous informe que seule la variable issue de l'outil Google Trends comporte une partie saisonnière. Nous procédons donc à sa désaisonnalisation en utilisant les méthodes X-13 et TRAMO-SEATS sous le logiciel JDemetra+.

➢ **TRAMO-SEATS**

TABLEAU N°12 : Désaisonnalisation par TRAMO-SEATS

• Le modèle

Summary

Estimation span: [I-2004 - III-2019]
63 observations
Series has been log-transformed
No trading days effects
No easter effect
1 detected outlier

Final model

Likelihood statistics
Number of effective observations = 58
Number of estimated parameters = 5

Loglikelihood = 84.30164987792381
Transformation adjustment = -235.65682888525978
Adjusted loglikelihood = -151.35517900733598

Standard error of the regression (ML estimate) = 0.055742879898953795
AIC = 312.71035801467195
AICC = 313.8642041685181
BIC (corrected for length) = -5.493980630778748

Scores at the solution
-0,003056 0,002731 0,001219

Arima model
[(1,1,1)(0,1,1)]

	Coefficients	T-Stat	P[|T| > t]
Phi(1)	-0,6394	-3,11	0,0030
Theta(1)	-0,1710	-0,66	0,5132
BTheta(1)	-0,5694	-4,96	0,0000

• Les conditions de validité

summary
Good

basic checks
definition: Good (0,000)
annual totals: Good (0,002)

regarima residuals
normality: Good (0,491)
independence: Good (0,971)
spectral td peaks: Good (0,157)
spectral seas peaks: Good (0,196)

outliers
number of outliers: Good (0,016)

seats
seas variance: Good (0,604)
irregular variance: Good (0,512)
seas/irr cross-correlation: Good (0,741)

residual seasonality tests
qs test on sa: Good (1,000)
f-test on sa (seasonal dummies): Good (0,861)
qs test on i: Good (1,000)
f-test on i (seasonal dummies): Good (0,871)

residual trading days tests
f-test on sa (td): Good (0,814)
f-test on i (td): Uncertain (0,043)

Aussi la méthode TRAMO-SEATS estime un modèle SARIMA(1,1,1)(0,1,1 $)_4$ avec une différenciation à la fois sur la partie classique ARMA et sur la partie saisonnière SMA. Seul le coefficient $\widehat{\theta_1}$ n'est pas significatif. Les résidus suivant un processus bruit blanc et une distribution normale, les conditions de validité fondamentales sont validées - il reste cependant une condition concernant la validité du calendrier de la composante irrégulière, qui est acceptée au seuil de 1% mais pas de 5%.

➢ **X-13**

TABLEAU N°13 : Désaisonnalisation par X-13

• Le modèle

Summary

Estimation span: [I-2004 - III-2019]
63 observations
Trading days effects (2 variables)
No easter effect

Final model

Likelihood statistics
Number of effective observations = 58
Number of estimated parameters = 5

Loglikelihood = -157.95493297545264
Standard error of the regression (ML estimate) = 3.630973798902074
AIC = 325.90986595090527
AICC = 327.0637121047514
BIC (corrected for length) = 2.8590323053729056

Scores at the solution
-0,000655 -0,00316

Arima model
[(0,1,1)(0,1,1)]

	Coefficients	T-Stat	P[|T| > t]
Theta(1)	0,4503	3,71	0,0005
BTheta(1)	-0,5527	-4,87	0,0000

• Les conditions de validité

summary
Good

basic checks
definition: Good (0,000)
annual totals: Good (0,008)

regarima residuals
normality: Good (0,340)
independence: Good (0,118)
spectral td peaks: Uncertain (0,062)
spectral seas peaks: Good (0,313)

outliers
number of outliers: Good (0,000)

m-statistics
q: Good (0,347)
q-m2: Good (0,388)

residual seasonality tests
qs test on sa: Good (1,000)
f-test on sa (seasonal dummies): Good (0,857)
qs test on i: Good (1,000)
f-test on i (seasonal dummies): Good (0,965)

residual trading days tests
f-test on sa (td): Good (0,727)
f-test on i (td): Good (0,781)

La méthode X-13 modélise, quant à elle, un modèle SARIMA(0,1,1)(0,1,1 $)_4$ - c'est-à-dire avec un processus moyenne mobile dans les parties classique et saisonnière. Les coefficients estimés sont tous deux significatifs et les conditions sont ici aussi, toutes validées sauf une qui reste secondaire dans la validation du modèle (elle concerne l'analyse spectrale des pics de fréquence des résidus). Comparons à présent ces 2 approches afin de déterminer quelle désaisonnalisation correspond le mieux à notre variable Google Trends.

TABLEAU N°14 : Comparaison des méthodes de désaisonnalisation

	TRAMO-SEATS	**X-13**
Modèle estimé	SARIMA(1,1,1)(0,1,1)$_4$	SARIMA(0,1,1)(0,1,1)$_4$
Nombre de paramètres significatifs	2/3	2/2
AICc*	313,86	327,06
Écart-type de régression	0,06	3,63
P-value normalité des résidus	0,49	0,34
P-value blancheur des résidus	0,97	0,19

* : nous regardons le critère d'Akaike corrigé puisque les modèles que nous comparons n'ont pas le même nombre de paramètres.

À partir du tableau n°14, on voit que les modèles issus des méthodes TRAMO-SEATS et X-13 sont très proches. On constate néanmoins que la qualité du modèle de SEATS est meilleure que celle de X-13 puisque l'écart type de régression et le critère d'Akaike sont plus faibles. De plus, les hypothèses nulles des tests de blancheur et de normalité des résidus sont plus largement acceptées que pour le modèle X-13. Ainsi, bien que le modèle SARIMA(1,1,1)(0,1,1)$_4$ ait un paramètre non significatif, nous préférons récupérer les données CVS de X_1 rendues non-saisonnières par ce modèle car il est bien spécifié et de bonne qualité. Nous importons à présent la variable Google Trends, ajustée des variations saisonnières par le logiciel JDemetra+, sous Gretl pour chercher à la cointégrer au taux de chômage, puis sous R pour les modélisations ARX. Pour finir, le graphique des séries initiale et CVS est disponible en annexe n°8 ; on voit comment les données sont lissées de tout comportement saisonnier.

B) Cointégration de la popularité du mot 'emploi' (X_1)

Nous rappelons que la première étape du processus vise à vérifier que l'ordre d'intégration est le même pour les 2 variables, c'est-à-dire I(1). Si ce n'est pas le cas, notre analyse s'arrête directement puisqu'il n'est pas possible de cointégrer des variables qui n'ont pas le même degré de différenciation. Nous procédons donc au test ADF sur la variable X_1 en niveau (mais désaisonnalisée) de manière à voir si celle-ci est stationnaire.

TABLEAU N°15 : Résultats du test ADF pour la série X_1 CVS

```
Test de Dickey-Fuller augmenté pour Popularite_emploi_CVS
testing down from 10 lags, criterion AIC
taille de l'échantillon 58
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 4 retards de (1-L)Popularite_emploi_CVS
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): 0,000145938
  statistique de test: tau_nc(1) = 0,025919
  p. critique asymptotique 0,6911
  Coeff. d'autocorrélation du 1er ordre pour e: 0,041
  différences retardées: F(4, 53) = 8,954 [0,0000]
```

On voit d'après le tableau n°15 que la p-value associée au test ADF est supérieure à 0,05 donc nous acceptons l'hypothèse nulle ; tout comme le taux de chômage, la variable issue de Google Trends n'est initialement pas stationnaire. Voyons à présent si elle l'est avec une différenciation, de manière à pouvoir cointégrer les 2 variables.

TABLEAU N°16 : Résultats du test ADF pour la série X_1 différenciée 1 fois

```
Test de Dickey-Fuller augmenté pour d_Popularite_emploi_CVS
testing down from 10 lags, criterion AIC
taille de l'échantillon 58
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 3 retards de (1-L)d_Popularite_emploi_CVS
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,392182
  statistique de test: tau_nc(1) = -2,92988
  p. critique asymptotique 0,003302
  Coeff. d'autocorrélation du 1er ordre pour e: 0,041
  différences retardées: F(3, 54) = 2,629 [0,0594]
```

D'après le tableau n°16 on remarque que la série différenciée une fois est effectivement stationnaire puisque la p-value du test ADF est inférieure au seuil de risque de 5%. Nous pouvons ainsi continuer la cointégration entre ces 2 variables car elles respectent la première condition du processus de cointégration. Pour cela nous construisons la relation de long terme en prenant les séries brutes, relation de laquelle nous sauvegardons les résidus de manière à vérifier leur stationnarité (première condition d'une relation non fallacieuse).

TABLEAU N°17 : Modèle de long terme et test ADF des résidus entre Y_t et X_1

```
Modèle 1: MCO, utilisant les observations 2004:1-2019:3 (T = 63)
Variable dépendante: Taux_chomage

                        coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------------------
  const                  6,90438      0,184863        37,35         9,85e-44    ***
  Popularite_emplo~      0,0338284    0,00282690      11,97         1,20e-17    ***

Moy. var. dép.          8,981320   Éc. type var. dép.      0,916827
Somme carrés résidus    15,56829   Éc. type de régression  0,505191
R2                      0,701273   R2 ajusté               0,696376
F(1, 61)                143,2001   p. critique (F)         1,20e-17
Log de vraisemblance   -45,35931   Critère d'Akaike        94,71863
Critère de Schwarz      99,00490   Hannan-Quinn            96,40444
rho                     0,906937   Durbin-Watson           0,190918
```

```
Test de Dickey-Fuller augmenté pour uhat1
testing down from 10 lags, criterion AIC
taille de l'échantillon 59
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 3 retards de (1-L)uhat1
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,162085
  statistique de test: tau_nc(1) = -2,72553
  p. critique asymptotique 0,006234
  Coeff. d'autocorrélation du 1er ordre pour e: 0,028
  différences retardées: F(3, 55) = 2,656 [0,0574]
```

Ainsi, d'après le tableau n°17 on constate que les résidus de long terme sont bien stationnaires car la p-value du test ADF s'élève à 0,006 ce qui est inférieur à 0,05 et nous

permet de rejeter l'hypothèse nulle de non-stationnarité des résidus. Nous avons donc $\widehat{\varepsilon_t} \sim I(0)$ qui nous permet de continuer notre démarche de cointégration de la variable issue de l'outil 'Google Trends' avec le taux de chômage français, en estimant cette fois la relation de court terme, c'est-à-dire avec les variables différenciées une fois auxquelles nous ajoutons les résidus de long terme retardés d'une période.

TABLEAU N°18 : Modèle de court terme (MCE) entre Y_t et X_1

```
Modèle 3: MCO, utilisant les observations 2004:2-2019:3 (T = 62)
Variable dépendante: d_Taux_chomage

                      coefficient  erreur std.  t de Student  p. critique
  ---------------------------------------------------------------------------
  d_Popularite_emp~    0,0261725   0,00862187       3,036       0,0035    ***
  uhat1_1             -0,110730    0,0585369       -1,892       0,0634    *

Moy. var. dép.          0,000482   Éc. type var. dép.      0,245735
Somme carrés résidus    2,801992   Éc. type de régression  0,216102
R2 non-centré           0,239321   R2 centré               0,239318
F(2, 60)                9,438436   p. critique (F)         0,000273
Log de vraisemblance    8,026723   Critère d'Akaike       -12,05345
Critère de Schwarz     -7,799178   Hannan-Quinn           -10,38311
rho                     0,131847   Durbin-Watson           1,730767
```

À partir du tableau n°18 qui nous donne les informations du modèle à correction d'erreur, on voit que le coefficient de correction est significatif (à 10%) et négatif puisque $\widehat{\delta}$ =-0,111. Ainsi la relation qui lie la popularité du mot 'emploi' au taux de chômage n'est pas fallacieuse ; en effet les résidus de long terme sont stationnaires et Y_t retourne bien à son équilibre de long terme. Aussi, la relation de court terme des variables 'taux de chômage' et 'popularité du mot emploi' nous informe que chaque trimestre, les 2 séries convergent de 11% environ - elles s'apparentent donc à 100% au bout de 9 périodes soit un peu plus de 2 ans.

Il est possible d'interpréter la relation de long terme qui lie Y_t et X_1 donnée en tableau n°17. Aussi nous voyons que le coefficient estimé de X_1 est significatif, l'outil de Google Trends est donc pertinent pour expliquer les variations du taux de chômage. Il est positif ce qui confirme la relation théorique et empirique qui lie ces 2 variables et que nous avons pu analyser en partie I à l'aide des graphiques de corrélation. Par conséquent, lorsque la popularité du mot 'emploi' augmente d'un point, le taux de chômage augmentera de 0,034%. La cointégration de la première variable a été concluante, voyons ce qu'il en est des 4 autres.

C) Cointégration du taux d'intérêt (X_2)

Nous procédons à la même analyse pour la deuxième variable explicative du taux de chômage ; le taux d'intérêt. Les résultats du test ADF, disponibles en tableau n°19 ci-dessous, indiquent que la série brute n'est pas stationnaire, en revanche elle l'est après une différenciation puisque la p-value est de 0 ce qui nous permet de rejeter H_0.

TABLEAU N°19 : Résultats du test ADF pour la série X_2 brute et différenciée 1 fois

```
Test de Dickey-Fuller augmenté pour Taux_interet
testing down from 20 lags, criterion AIC
taille de l'échantillon 60
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 2 retards de (1-L)Taux_interet
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,0210602
  statistique de test: tau_nc(1) = -1,74787
  p. critique asymptotique 0,07643
  Coeff. d'autocorrélation du 1er ordre pour e: 0,015
  différences retardées: F(2, 57) = 4,040 [0,0229]
```

```
Test de Dickey-Fuller augmenté pour d_Taux_interet
testing down from 20 lags, criterion AIC
taille de l'échantillon 61
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 0 retards de (1-L)d_Taux_interet
  modèle: (1-L)y = (a-1)*y(-1) + e
  valeur estimée de (a - 1): -0,740098
  statistique de test: tau_nc(1) = -5,80797
  p. critique 3,243e-08
  Coeff. d'autocorrélation du 1er ordre pour e: 0,073
```

Étant donné que les 2 séries sont intégrées du même ordre, nous pouvons passer à l'analyse des résidus de long terme en commençant par estimer un modèle en régression linéaire simple composé seulement de ces 2 séries, comme pour la cointégration précédente. La sortie du modèle sous Gretl est disponible en annexe n°9, après sauvegarde des résidus on teste leur stationnarité à l'aide du test KPSS. Également disponible en annexe n°9 la sortie de ce test indique que les résidus associés au modèle de long terme sont stationnaires puisque la valeur de la statistique "p" est supérieure au seuil de significativité fixé à 5%. Ici encore, nous pouvons poursuivre les démarches de cointégration du taux d'intérêt avec le taux de chômage.

TABLEAU N°20 : Modèle de court terme (MCE) entre Y_t et X_2

```
Modèle 8: MCO, utilisant les observations 2004:2-2019:3 (T = 62)
Variable dépendante: d_Taux_chomage

                  coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------------------
  const           -0,00680439   0,0325114     -0,2093        0,8349
  d_Taux_interet  -0,112544     0,117268      -0,9597        0,3411
  uhat7_1         -0,0219267    0,0508102     -0,4315        0,6676

Moy. var. dép.          0,000482   Éc. type var. dép.      0,245735
Somme carrés résidus    3,618237   Éc. type de régression  0,247641
R2                      0,017724   R2 ajusté              -0,015573
F(2, 59)                0,532308   p. critique (F)         0,590042
Log de vraisemblance    0,101383   Critère d'Akaike        5,797234
Critère de Schwarz      12,17864   Hannan-Quinn            8,302734
rho                     0,228340   Durbin-Watson           1,540318
```

La relation entre le taux de chômage et le taux d'intérêt est fallacieuse, elle n'est pas pertinente puisque le coefficient de correction associé au modèle de correction d'erreur est non significatif (confère tableau n°20). Nous ne poursuivons donc pas les démarches de cointégration entre ces 2 variables.

D) Cointégration de la production industrielle (X_3)

Les résultats des tests pour la variable “production industrielle” sont les mêmes que ceux de la variable Google Trends ; X_3 est intégré d’ordre 1 comme le taux de chômage, il est par conséquent possible de commencer les démarches de cointégration en passant par la régression linéaire simple. Comme pour les variables X_1 et X_2 les résidus sont stationnaires, c'est-à-dire intégrés d’ordre 0 car la p-value du test KPSS est supérieure à α=0,05. Enfin, lorsque l’on estime la relation de court terme par le MCE il apparaît que delta estimé est négatif et significatif, nous sommes dans le cas d’une relation non fallacieuse qu’il est donc possible d’interpréter. Le détail de cette analyse est disponible en annexe n°10. On voit ainsi que les séries du taux de chômage et de la production industrielle, qui reflète l’activité économique, convergent de 11% par trimestre ($\widehat{\delta}$=-0,1089). Elles s’apparentent donc totalement au bout de 10 périodes ($\frac{100}{10,89}$ = 9,18) ce qui correspond à 2 ans et demi. Aussi, le coefficient estimé de X_3 à long terme s’élève à −0,106 : lorsque la production industrielle augmente d’une unité par rapport à la base 100 en 2015, le taux de chômage diminue de 0,11 point de pourcentage.

E) Cointégration de la population active (X_4)

Nous procédons à la cointégration de X_4 avec le taux de chômage. D’après le tableau n°21 on voit que les 2 séries ont le même ordre d’intégration d=1 puisque X_4 est stationnaire lorsqu’il est différencié une fois. De plus, on voit que les résidus sauvegardés de la régression de long terme entre les 2 séries sont stationnaires, puisque la p-value du test ADF est de 0,022 ce qui nous permet de rejeter l’hypothèse nulle selon laquelle les résidus ne sont pas stationnaires.

TABLEAU N°21 : Développement démarche de cointégration de Y_t avec la population active

```
Test de Dickey-Fuller augmenté pour Population_active
testing down from 20 lags, criterion AIC
taille de l'échantillon 62
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 0 retards de (1-L)Population_active
  modèle: (1-L)y = (a-1)*y(-1) + e
  valeur estimée de (a - 1): 0,00141609
  statistique de test: tau_nc(1) = 2,71892
  p. critique 0,9982
  Coeff. d'autocorrélation du 1er ordre pour e: -0,212
```

```
Test de Dickey-Fuller augmenté pour d_Population_active
testing down from 20 lags, criterion AIC
taille de l'échantillon 61
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 0 retards de (1-L)d_Population_active
  modèle: (1-L)y = (a-1)*y(-1) + e
  valeur estimée de (a - 1): -1,06963
  statistique de test: tau_nc(1) = -8,25518
  p. critique 7,336e-21
  Coeff. d'autocorrélation du 1er ordre pour e: 0,007
```

```
Modèle 19: MCO, utilisant les observations 2004:1-2019:3 (T = 63)
Variable dépendante: Taux_chomage

                     coefficient   erreur std.  t de Student  p. critique
  ---------------------------------------------------------------------------
  const              -9,42918      2,84188        -3,318      0,0015     ***
  Population_active   0,000645274  9,95562e-05     6,482      1,79e-08   ***

Moy. var. dép.          8,981320   Éc. type var. dép.      0,916827
Somme carrés résidus    30,86155   Éc. type de régression  0,711285
R2                      0,407824   R2 ajusté               0,398116
F(1, 61)                42,00990   p. critique (F)         1,79e-08
Log de vraisemblance   -66,91398   Critère d'Akaike        137,8280
Critère de Schwarz      142,1142   Hannan-Quinn            139,5138
rho                     0,960119   Durbin-Watson           0,106989
```

```
Test de Dickey-Fuller augmenté pour uhat19
testing down from 20 lags, criterion AIC
taille de l'échantillon 60
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 2 retards de (1-L)uhat19
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,0948566
  statistique de test: tau_nc(1) = -2,2761
  p. critique asymptotique 0,02205
  Coeff. d'autocorrélation du 1er ordre pour e: -0,037
  différences retardées: F(2, 57) = 7,810 [0,0010]
```

Le fait que les résidus soient intégrés d'ordre 0 nous permet de valider la première condition pour que la relation entre Y_t et X_4 ne soit pas fallacieuse, on peut maintenant construire le modèle à correction d'erreur noté MCE qui modélise la relation de court terme qu'il existe entre le taux de chômage et la population active.

TABLEAU N°22 : MCE pour Y_t et X_4

```
Modèle 20: MCO, utilisant les observations 2004:2-2019:3 (T = 62)
Variable dépendante: d_Taux_chomage

                      coefficient   erreur std.  t de Student  p. critique
  ---------------------------------------------------------------------------
  const               -0,0293921    0,0314125      -0,9357       0,3533
  d_Population_act~    0,000746067  0,000257412     2,898        0,0053     ***
  uhat19_1            -0,0409446    0,0431239      -0,9495       0,3463

Moy. var. dép.          0,000482   Éc. type var. dép.      0,245735
Somme carrés résidus    3,206684   Éc. type de régression  0,233132
R2                      0,129453   R2 ajusté               0,099942
F(2, 59)                4,386723   p. critique (F)         0,016745
Log de vraisemblance    3,844620   Critère d'Akaike       -1,689240
Critère de Schwarz      4,692163   Hannan-Quinn            0,816260
rho                     0,323455   Durbin-Watson           1,347502
```

Le tableau n°22 contient l'estimation du modèle à correction d'erreur de court terme englobant les résidus de long terme retardés d'une période. On voit que le coefficient de correction est bien négatif puisque $\widehat{\delta}$ =-0,041, en revanche il n'est pas significatif donc nous ne pouvons pas interpréter les résultats de la régression linéaire de long terme car cela implique une relation fallacieuse entre Y_t et X_4.

Par conséquent, les variables 'taux de chômage' et 'population active' sont cointégrées (avec d=1 et $\widehat{\varepsilon}_t \sim I(0)$) mais leur relation est fallacieuse donc la cointégration de ces 2 séries n'est pas concluante. Regardons, pour terminer, la nature de la relation qui lie les effectifs de l'industrie au taux de chômage, si elle n'est pas fallacieuse nous pourrons nous servir de cette variable pour modéliser et prévoir le taux de chômage.

F) Cointégration des effectifs dans l'industrie (X_5)

La relation entre Y_t et X_5 n'est pas fallacieuse. En effet, les variables sont intégrées du même ordre à savoir d=1, les résidus de la relation à long terme sont stationnaires et le coefficient de correction est négatif et significatif. On peut donc affirmer à partir des résultats des tests disponibles en annexe n°11 que ces deux variables sont cointégrées. $\widehat{\delta}$ =-0,052 signifie que les séries convergent l'une vers l'autre de 5% chaque trimestre ; ainsi le point de convergence se situe à 19 périodes soit près de 5 ans.

Pour finir, le tableau n°23 résume les démarches de cointégration des 5 variables explicatives avec le taux de chômage, dont 3 ont été concluantes. Les 2 autres variables n'étaient pas 'cointégrablent' au taux de chômage parce que les conditions n'étaient pas vérifiées, *i.e.* même ordre d'intégration des 2 séries et relation non fallacieuse. Les intuitions que nous avions eues en section I de cette partie se sont avérées justes puisqu'effectivement, les 3 variables qui suivaient les mêmes évolutions que le taux de chômage sont finalement cointégrées à ce dernier.

TABLEAU N°23 : Récapitulatif des processus de cointégration des variables

Variables	Ordre d'intégration	Stationnarité des résidus	Coefficient de correction CT	Conclusion
Popularité du mot emploi et Y_t	d=1	p-value test ADF=0,029	$\widehat{\delta}$ =-0,041 et significatif	✓
Taux d'intérêt et Y_t	d=1	p-value test KPSS>0,05	$\widehat{\delta}$ =-0,022 mais non significatif	✗
Production industrielle et Y_t	d=1	p-value test KPSS>0,05	$\widehat{\delta}$ =-0,109 et significatif	✓
Population active et Y_t	d=1	p-value test ADF=0,0221	$\widehat{\delta}$ =-0,041 mais non significatif	✗
Effectifs industrie et Y_t	d=1	p-value test KPSS>0,1	$\widehat{\delta}$ =-0,052 et significatif	✓

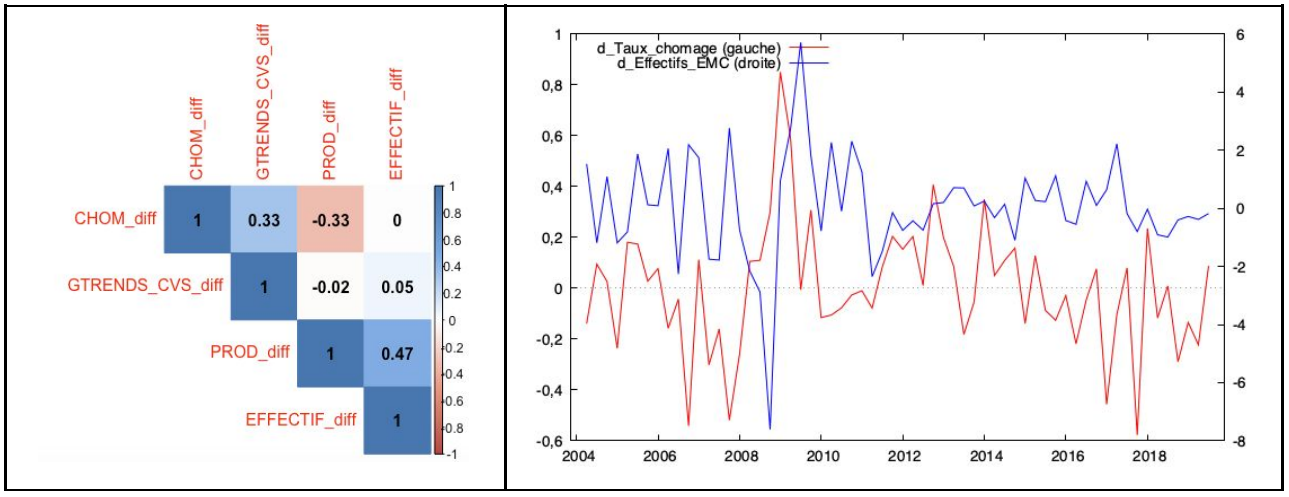
: test KPSS (H_0 : variable stationnaire)

: test ADF (H_0 : variable non stationnaire)

IV. Modélisations et prévisions ARX

Nous savons désormais que parmi les 5 variables explicatives que nous avions sélectionnées pour analyser Y_t, seules 3 d'entre elles ont une relation stable et robuste. Les autres variables n'étant pas cointégrées au taux de chômage, s'en servir comme variable explicative n'aurait pas de sens puisqu'elles ont une relation biaisée avec celui-ci. Sous R studio nous avons donc différencié les 3 variables cointégrées que nous avons rassemblées en une base appelée "df_CVS_diff" composée du taux de chômage différencié ainsi que des 3 variables. Avant d'entreprendre la recherche de modèles et la prévision à partir de ces derniers, nous nous intéressons aux corrélations qui lient les variables différenciées entre elles.

FIGURE N°3 : Matrice de corrélation des variables différenciées et graphique d'évolution



À partir de la figure n°3, on voit d'emblée qu'aucune relation ne lie le taux de chômage aux effectifs de l'industrie qui, nous le rappelons, sont issus des enquêtes mensuelles de conjoncture réalisées par la Banque de France. Le fait que le coefficient de corrélation soit égal à zéro est étrange puisqu'il signifie que les 2 variables n'ont aucun caractère commun. Pourtant, lorsque l'on observe le graphique qui met les variables différenciées côte à côte, il apparaît qu'elles suivent globalement les mêmes évolutions, quoique l'une varie entre -0,577 et 0,85 (taux de chômage), et l'autre entre -7,61 et 5,71 (effectifs dans l'industrie). En outre, nous sommes contraints de ne pas utiliser cette variable car, bien que cointégrée au taux de chômage, il n'existe aucune relation qui la lie à Y_t ; elle ne peut donc pas servir à l'expliquer.

Par ailleurs, les variables Google Trends et production industrielle sont corrélées positivement et négativement avec un coefficient de 0,33. De plus, la variable google Trends

étant faiblement corrélée à la production industrielle (CC[43]=-0,02) nous n'aurons pas de problème de multicolinéarité et serons donc sereins lors de l'estimation des modèles.

A) Modélisations

Par conséquent dans cette partie nous estimerons 3 modèles différents :

- ARX1 : Y_i fonction de X_1 et X_3
- ARX2 : Y_i fonction de X_1
- ARX3 : Y_i fonction de X_3

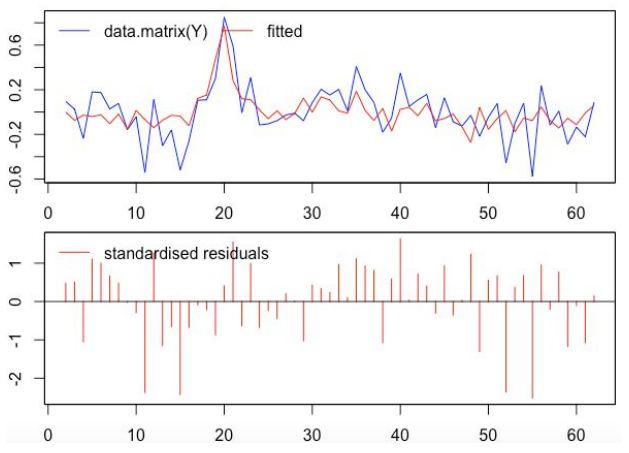
Où :

- Y_i correspond au taux de chômage différencié noté '**CHOM_diff**'
- X_1 correspond à la popularité du mot 'emploi' sous Google Trends, variable qui a été initialement désaisonnalisée, réimportée puis différenciée ; elle est donc notée '**GTRENDS_CVS_diff**'
- X_3 correspond à la production industrielle différenciée, notée '**PROD_diff**'
- i correspond à la période qui s'étend du 2è trimestre de 2004 au 3è de 2019

➢ ARX1

TABLEAU N°24 : Estimation du modèle ARX1

```
> ARX1_1

Date: Sat May 23 12:20:32 2020
Dependent var.: data.matrix(Y)
Method: Ordinary Least Squares (OLS)
Variance-Covariance: Ordinary
No. of observations (mean eq.): 61
Sample: 2 to 62

Mean equation:

                      coef  std.error  t-stat   p-value
mconst          -0.0111678  0.0256507 -0.4354 0.6649302
ar1              0.0851519  0.1137324  0.7487 0.4571143
GTRENDS_CVS_diff 0.0238450  0.0083301  2.8625 0.0058697
PROD_diff       -0.0549188  0.0134600 -4.0802 0.0001416

Diagnostics and fit:

                  Chi-sq df p-value
Ljung-Box AR(2)   2.28205  2 0.31949
Ljung-Box ARCH(1) 0.42666  1 0.51363

SE of regression  0.19907
R-squared         0.38345
Log-lik.(n=61)   13.90432
```



[43] **CC** : Coefficient de corrélation

On voit sur le graphique ci-dessus que les valeurs prédites par le modèle suivent de près les variations des valeurs observées ; composé de seulement 2 variables explicatives, ce premier modèle ARX prédit relativement bien le taux de chômage avec une qualité de modèle s'élevant à 38,35%. Cependant, la constante n'étant pas significative nous recommençons l'estimation en retirant cette dernière.

TABLEAU N°25 : Estimation du modèle ARX1 sans la constante

```
> ARX1_2

Date: Sat May 23 12:21:36 2020
Dependent var.: data.matrix(Y)
Method: Ordinary Least Squares (OLS)
Variance-Covariance: Ordinary
No. of observations (mean eq.): 61
Sample: 2 to 62

Mean equation:

                       coef  std.error  t-stat   p-value
ar1               0.0874126  0.1128172  0.7748 0.4415959
GTRENDS_CVS_diff  0.0235051  0.0082353  2.8542 0.0059741
PROD_diff        -0.0546521  0.0133518 -4.0932 0.0001333

Diagnostics and fit:

                    Chi-sq df p-value
Ljung-Box AR(2)   2.32823  2 0.31220
Ljung-Box ARCH(1) 0.48513  1 0.48611

SE of regression  0.19768
R-squared         0.38140
Log-lik.(n=61)   13.83351
```

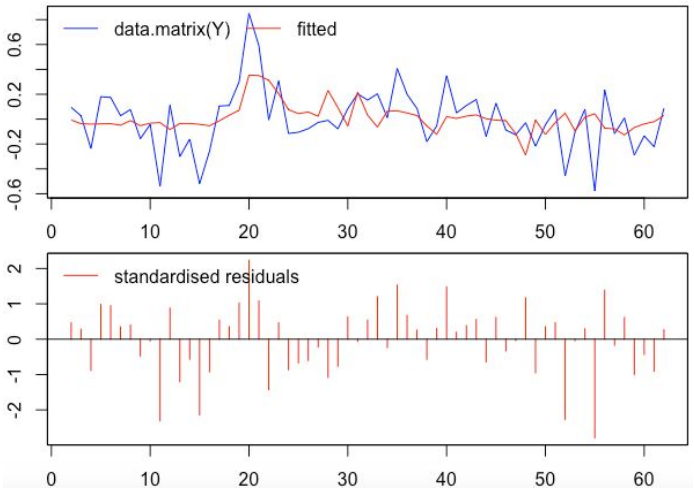
Le Log de vraisemblance de ce modèle s'élève à 13,83 et son écart-type à 0,198 : nous regarderons ces critères pour chacun des 3 modèles ARX construits, de manière à évaluer leur qualité en les comparant les uns aux autres. Nous passons maintenant à l'estimation d'un modèle ARX où seule la variable de Google Tendances est utilisée pour modéliser Y_t.

➢ ARX2

De même manière que la première estimation ARX, la constante de ce modèle n'était pas significative (confère annexe n°12). Nous l'avons donc estimé une seconde fois, les résultats de la modélisation sont disponibles ci-dessous :

TABLEAU N°26 : Estimation du modèle ARX2 sans la constante

```
> ARX2_2

Date: Mon May 25 10:59:31 2020
Dependent var.: data.matrix(Y)
Method: Ordinary Least Squares (OLS)
Variance-Covariance: Ordinary
No. of observations (mean eq.): 61
Sample: 2 to 62

Mean equation:

                      coef std.error t-stat  p-value
ar1              0.0894837 0.1269884 0.7047 0.483794
GTRENDS_CVS_diff 0.0291974 0.0091367 3.1956 0.002242

Diagnostics and fit:

                   Chi-sq df p-value
Ljung-Box AR(2)    1.4784  2 0.47750
Ljung-Box ARCH(1) 0.1852  1 0.66694

SE of regression 0.22251
R-squared        0.20270
Log-lik.(n=61)   6.11496
```



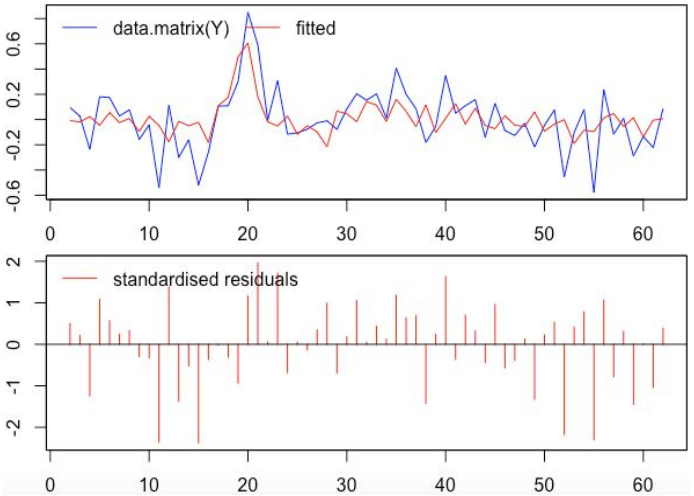
Étant donné que la seule variable exogène est la popularité du mot 'emploi', on voit que les valeurs prédites par le modèle suivent moins les variations de Y_t observé, parce que l'information est plus faible par rapport au premier modèle ARX. Cela se confirme par un écart-type de régression plus élevé, et un logarithme de vraisemblance ainsi qu'un coefficient de détermination plus faible. Nous estimons un dernier modèle ARX, composé cette fois d'une seule variable explicative qu'est la production industrielle, notée X_3.

➢ **ARX3**

Encore une fois la constante n'était pas significative, nous avons donc réestimé le modèle que voilà ci-dessous :

TABLEAU N°27 : Estimation du modèle ARX3 sans la constante

```
> ARX3_2

Date: Mon May 25 11:01:35 2020
Dependent var.: data.matrix(Y)
Method: Ordinary Least Squares (OLS)
Variance-Covariance: Ordinary
No. of observations (mean eq.): 61
Sample: 2 to 62

Mean equation:

                 coef std.error  t-stat   p-value
ar1          0.215994  0.109517  1.9722   0.05328
PROD_diff -0.061087  0.013934 -4.3840 4.866e-05

Diagnostics and fit:

                   Chi-sq df p-value
Ljung-Box AR(2)   3.61194  2 0.16432
Ljung-Box ARCH(1) 0.77335  1 0.37918

SE of regression 0.20931
R-squared        0.29451
Log-lik.(n=61)   9.84635
```



On constate cette fois, que le paramètre ar1 est lui aussi significatif au seuil de risque $\alpha = 0,1$ contrairement aux modèles précédents. De plus, le modèle semble bien prédire les fluctuations de Y_t comme visible sur le graphique de droite, et le coefficient de détermination mesuré par R^2 est plus élevé que dans le cas du modèle ARX2 qui contenait seulement la variable Google Trends. Nous pouvons donc résumer l'information de ces 3 modélisations dans un tableau récapitulatif ; celui-ci permet de constater les points forts et faibles de chaque modèle avant d'entreprendre les prévisions.

TABLEAU N°28 : Récapitulatif des modèles ARX estimés

	ARX1	**ARX2**	**ARX3**
Variables explicatives	X_1 et X_3	X_1	X_3
Nombre de coefficients significatifs à 10%	1/2	1/2	2/2
Écart-type de régression	0,20	0,22	0,21
R^2	0,38	0,20	0,29
Log de vraisemblance	13,83	6,11	9,85

D'après le tableau n°28, on voit que le meilleur modèle au vu des critères présentés dans la régression, est celui qui comporte à la fois la variable de la production industrielle X_3 connue pour son apport dans les prévisions du taux de chômage, et X_1 comme variable complémentaire permettant d'améliorer la qualité du modèle (le log de vraisemblance ainsi que le coefficient de détermination sont plus élevés dans le modèle ARX1 que dans le modèle ARX3 qui ne prend pas en compte la variable Google Trends). En outre, à la seule observation des modèles, on voit que la variable issue de l'outil Google Trends permet d'améliorer la vraisemblance, de réduire l'écart-type mais aussi d'augmenter le coefficient de détermination qui mesure la qualité de prédiction d'une régression linéaire (ARX1 vs. ARX3). Cependant, lorsque l'outil est pris comme seul variable explicative du taux de chômage, on voit que son apport est moins notable que celui de variables traditionnelles comme la production industrielle. Le coefficient R^2 et le log de vraisemblance sont ainsi plus faibles dans le modèle n'incluant que la popularité du mot 'emploi', par rapport au modèle n'incluant que la production industrielle (ARX2 vs. ARX3).

Nous réalisons à présent les prévisions à partir de ces 3 différents modèles aux horizons h=1, h=4 et h=8. En effet, nous nous intéressons dans ce travail à l'apport de Google Tendances pour prévoir le chômage et l'emploi : en réalisant des prévisions à partir de différents modèles incluant ou n'incluant pas ce dernier, nous pourrons nous rendre compte de sa contribution à la précision des prévisions à plusieurs horizons.

B) Prévisions

Les valeurs des prévisions du taux de chômage différencié, aux 3 horizons et par les modèles estimés dans la section précédente, sont visibles en annexe n°13. On constate que les prévisions restent relativement différentes entre les modèles, que ce soit à court, moyen, ou long terme. Comme expliqué précédemment, nous procéderons à une transformation de ces prévisions pour les remettre au niveau de la série initiale.

V. Comparaison des prévisions

Vient à présent la partie la plus intéressante du mémoire, celle où nous allons pouvoir réellement répondre à la problématique en comparant nos différentes prévisions. Pour rappel nous disposons de prévisions à 3 horizons différents (h=1, h=4 et h=8) et de 4 différents modèles : un modèle AR(1) trouvé par l'application de la méthodologie qui, par sa simplicité, nous servira de référence par rapport aux 3 modèles ARX qui incluent, eux, des variables explicatives.

A) Transformation des prévisions

Nous commençons par transformer les prévisions de la manière suivante : pour une prévision qui s'étend de T à T+4, nous prenons la valeur du taux de chômage en niveau en T-1, à laquelle nous additionnons la prévision pour la période T. À cette valeur "en niveau" nous additionnons la prévision de Y_t à la période T+1, nous répétons l'opération jusqu'à atteindre l'horizon de prévision. De cette manière, nous nous servons de la prévision calculée en T+1 pour prédire le taux de chômage en T+2 etc. Nous obtenons les prévisions suivantes :

TABLEAU N°29 : Prévisions transformées à horizon d'une période

	ARIMA	ARX1	ARX2	ARX3
T3 2019	8.41986	8.517972	8.48019	8.4602

TABLEAU N°30 : Prévisions transformées à horizon de 4 périodes

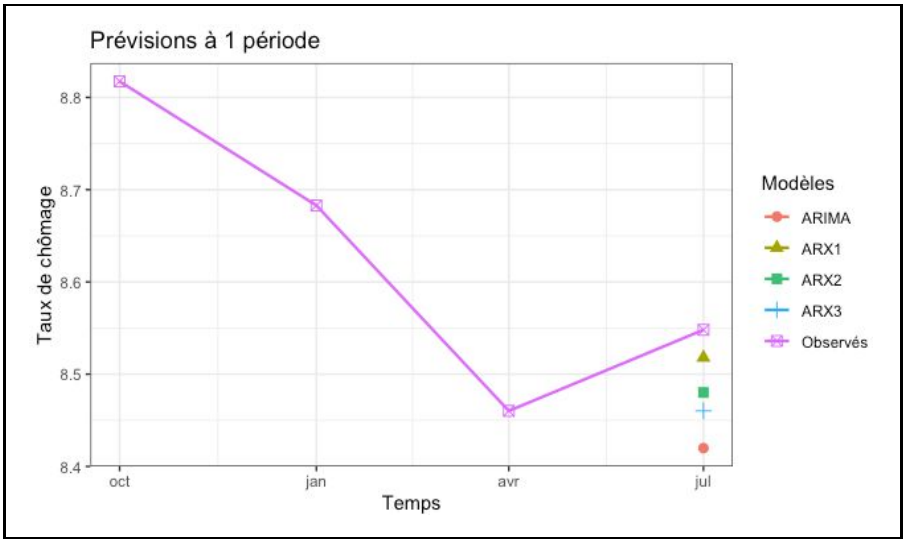
	ARIMA	ARX1	ARX2	ARX3
T4 2018	9.116076	8.969417	8.979848	9.052820
T1 2019	9.125377	8.924366	8.915995	9.069321
T2 2019	9.134682	8.817408	8.878422	8.930930
T3 2019	9.143987	8.812992	8.854208	8.923648

TABLEAU N°31 : Prévisions transformées à horizon de 8 périodes

	ARIMA	ARX1	ARX2	ARX3
T4 2017	9.595703	9.509883	9.565211	9.478112
T1 2018	9.621309	9.462046	9.618032	9.411702
T2 2018	9.642120	9.481774	9.501768	9.385653
T3 2018	9.661279	9.415089	9.423319	9.440174
T4 2018	9.679870	9.277714	9.296376	9.386202
T1 2019	9.698264	9.232664	9.232523	9.402704
T2 2019	9.716590	9.125706	9.194951	9.264313
T3 2019	9.734893	9.121289	9.170736	9.257031

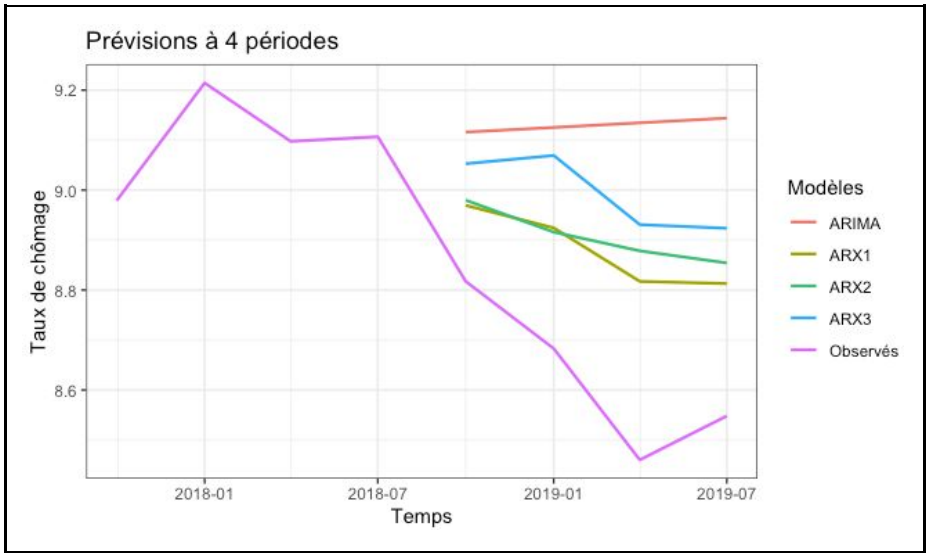
B) Représentations graphiques des prévisions

GRAPHIQUE N°13 : Prévisions à h=1



D'après le graphique n°13, on voit que la prévision s'approchant le plus de la valeur observée du troisième trimestre de 2019, est celle issue du modèle ARX1 c'est-à-dire avec la popularité du mot 'emploi', et la production industrielle comme variables explicatives. Par ailleurs, la deuxième meilleure prévision est issue du modèle ARX2 contenant seulement la variable Google Trends. Ainsi, il semble qu'à très court terme (h=1) le modèle ARX2 prévoit mieux le chômage que le modèle ARX3 qui n'est constitué que de la production industrielle. La prévisions du modèle ARIMA, *i.e.* AR(1), est celle qui s'éloigne le plus de la valeur réelle du taux de chômage.

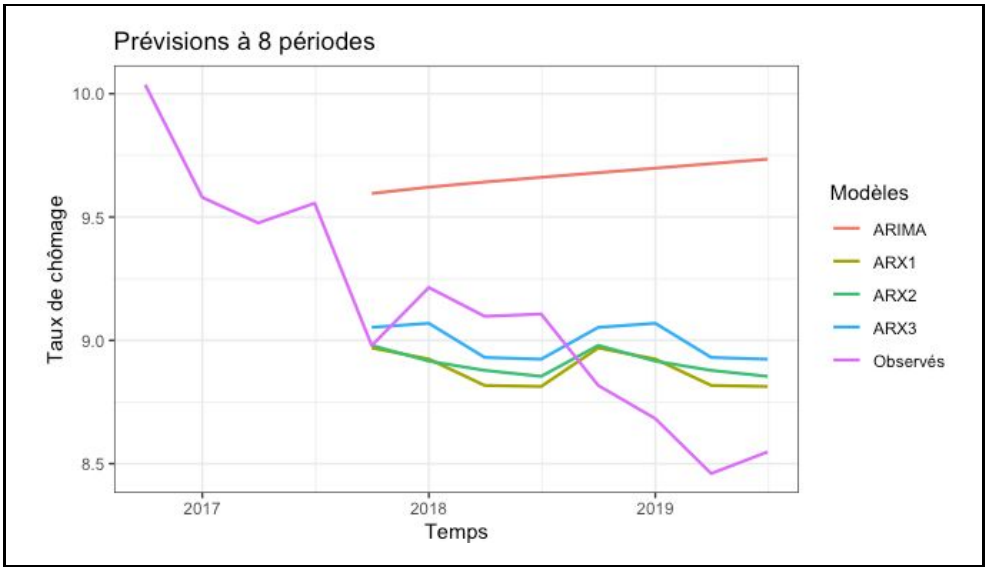
GRAPHIQUE N°14 : Prévisions à h=4



Comme nous l'avons expliqué dans la section précédente, nous avons transformé les prévisions du taux de chômage différencié en prenant la valeur la dernière valeur du taux de chômage en niveau avant que ne commencent les prévisions, valeur à laquelle nous avons additionné la première valeur de la prévision. C'est à partir de ce résultat que nous avons additionné la deuxième valeur de la prévision etc. Étant donné qu'à partir du troisième trimestre de l'année 2018, le taux de chômage a fortement baissé, les différents modèles ont eu du mal à prévoir une telle décroissance. On comprend donc que les erreurs des prévisions à 4 périodes seront plus importantes qu'à une période, ce qui semble assez logique.

Cependant, le modèle qui paraît le plus précis est le modèle ARX1 encore une fois, suivit de l'ARX2 - ce qui était le cas aussi pour les prévisions à horizon 1. Tandis que le taux de chômage diminue fortement à partir du 3è trimestre de 2018, le modèle benchmark prévoit une hausse constante de ce dernier, on imagine que les taux d'erreurs seront donc particulièrement élevés pour les prévisions ARIMA.

GRAPHIQUE N°15 : Prévisions à h=8



On voit d'emblée sur le graphique n°15 que les prévisions ARIMA sont extrêmement mauvaises. En effet, les erreurs augmentent avec l'horizon parce qu'à long terme la série tend vers sa moyenne, les prévisions ARIMA sont donc surtout utiles à court terme. Les modèles ARIMA ont une mémoire courte qui implique une grande vitesse de convergence expliquant ainsi les prévisions presque aberrantes de ce dernier.

Il est difficile de déterminer quelle prévision est meilleure qu'une autre au sein des modèles ARX puisqu'elles sont toutes les 3 en dessous des vraies valeurs du T3 2017 au T3 2018, puis au dessus jusqu'au 3è trimestre de 2019. C'est pour cela que nous nous intéressons maintenant aux erreurs de prévision de chacune d'entre elles.

C) Erreurs de prévisions

Nous allons dans cette section regarder les erreurs de prévisions des 4 modèles aux 3 horizons de prévision, en s'intéressant aux 4 types d'erreur énoncés précédemment :

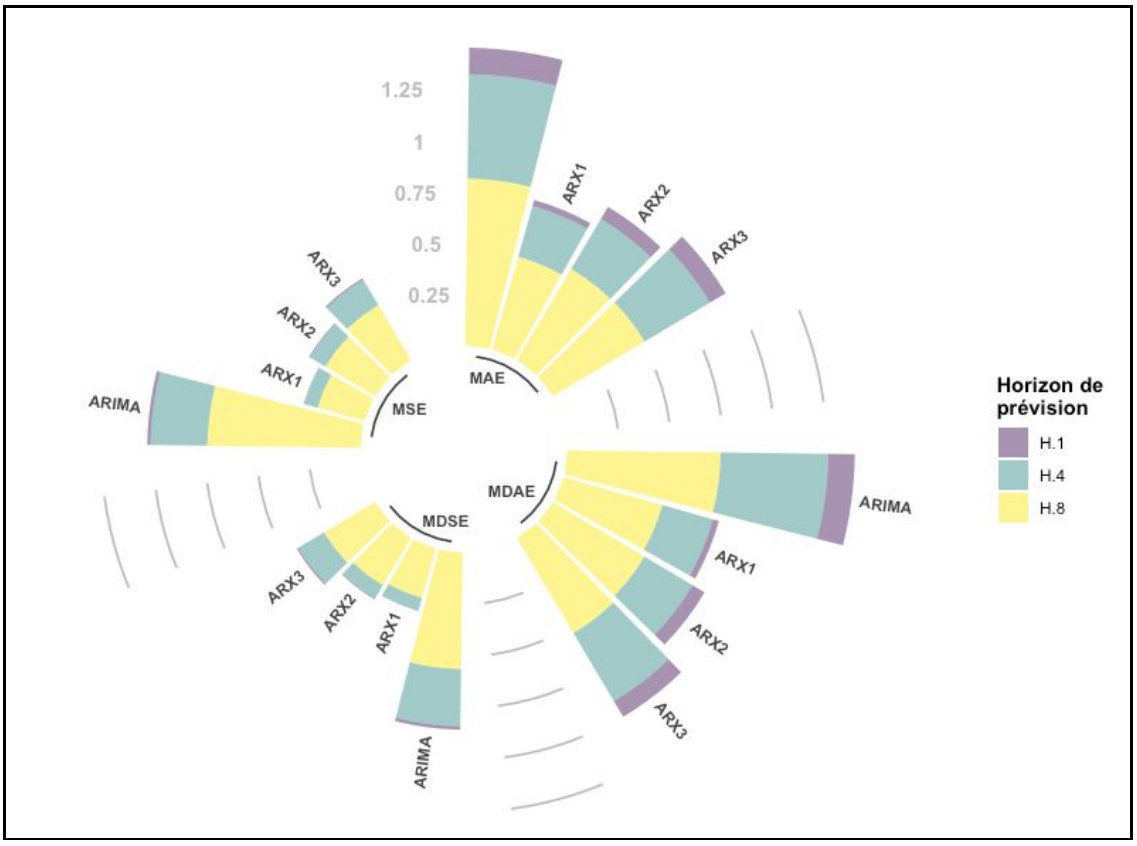
- l'erreur absolue moyenne : MAE *(Mean Absolute Error)*
- l'erreur quadratique moyenne : MSE *(Mean Squared Error)*
- l'erreur absolue médiane : MDAE *(MeDian Absolute Error)*
- l'erreur quadratique médiane : MDSE *(MeDian Squared Error)*

Nous représenterons chacune de ces erreurs sur un graphique en barplot circulaire, élaboré à partir du site "R Graph Gallery".[44] Nous étudierons ces dernières grâce au code couleur des

[44] **'The R Graph Gallery'** offre une collection de graphiques réalisables sous R, il se concentre surtout sur les packages *tidyverse* et *ggplot2*. (consulté le 25/05/2020)

horizons (jaune=H8, vert=H4 et violet=H1), et en nous appuyant sur les valeurs exactes des erreurs de prévisions disponibles en annexe n°14.

GRAPHIQUE N°16 : Barplot circulaire des erreurs de prévision



Comme nous l'avions pressenti en regardant les graphiques de prévisions dans la section précédente, il est clair que les erreurs croissent avec l'horizon de prévision. Plus on essaye de prévoir une variable loin, plus la probabilité de se tromper augmente. De même, ce sont les prévisions du modèle ARIMA qui sont toujours les moins bonnes puisque les taux d'erreur surpassent les autres, que ce soit l'erreur absolue ou quadratique, médiane ou moyenne.

➢ À horizon d'une période, il convient de regarder les erreurs absolues puisqu'on ne distingue pas clairement les différences de taux d'erreur entre les 4 modèles lorsque l'on regarde les erreurs quadratiques. On voit alors que le modèle ARIMA a les plus grands taux d'erreur, il est suivi du modèle ARX3 puis de l'ARX2 pour finir par l'ARX1 qui a, par conséquent, les taux d'erreur les plus faibles. Cela signifie qu'à court terme, le modèle qui prédit au mieux le taux de chômage est celui combinant à la fois la popularité du mot 'emploi' et la production industrielle. Par ailleurs, lorsque l'on regarde les modèles composés d'une seule des 2 variables explicatives énumérées ci-dessus, c'est le modèle ARX2 qui 'gagne'. Prise

individuellement, la variable de Google Trends prédit mieux le taux de chômage qu'une variable traditionnelle ayant pour objectif de refléter l'activité économique.

➢ À horizon de 4 périodes, le constat est le même c'est-à-dire que les prévisions réalisées à partir du modèle AR(1) contiennent les erreurs les plus grandes, et c'est le modèle combinant les 2 variables explicatives qui prédit le mieux Y_t suivit par le modèle ARX2 puis ARX3. La conclusion est donc similaire à celle pour un horizon de prévision d'une période.

➢ À horizon de 8 périodes on voit que les écarts de prévisions entre les modèles sont moins importants, proportionnellement au niveau des taux d'erreur, comme visible si l'on regarde les erreurs absolues médianes des 4 modèles. Ainsi, il semble que les meilleures prévisions soient issues du modèle ARX1. En revanche, il est plus difficile d'affirmer que le modèle ARX2 prévoit mieux Y_t que ARX3, étant donné que la différence entre leurs taux est de seulement 0,018 en moyenne sur les 4 types d'erreur (cf. annexe n°14).

De manière à pouvoir conclure sur la précision des prévisions, nous allons appliquer le test de Diebold-Mariano aux 3 horizons pour prouver statistiquement quel modèle est meilleur qu'un autre.

D) Tests de Diebold-Mariano

Nous travaillons pour cette section au seuil de significativité de 5%. Nous réalisons le test dans les 2 sens, puisque s'il n'y a pas de différence significative dans un sens, cela n'implique pas qu'il puisse ne pas y avoir dans l'autre sens.

TABLEAU N°32 : Test DM à horizon d'une période

	ARIMA	**ARX1**	**ARX2**	**ARX3**
ARIMA		0.8996	0.6964	0.8331
ARX1	0.1004		0.02121	0.04586
ARX2	0.3036	0.9788		0.825
ARX3	0.1669	0.9541	0.175	

D'après le tableau n°32, on constate que le modèle ARX1 est le meilleur puisqu'au mieux il n'y a pas de différence significative avec les autres modèles, sinon il a une capacité prédictionnelle supérieure comme nous le voyons quand il est comparé aux modèles ARX2 et ARX3 avec des p-values respectivement de 0,021 et 0,046. D'autre part, le test appliqué à H=1

ne révèle pas que le modèle ARX2 est plus précis que l'ARX3 contrairement à ce que l'on aurait pu penser au regard des graphiques de prévision et des taux d'erreur. Aussi, la probabilité du test DM étant de 0,1004, il apparaît qu'à 10% le modèle ARIMA n'est pas moins précis que le modèle ARX1 ou encore ARX2 et 3. Regardons ce qu'il en est pour les prévisions sur 4 trimestres.

TABLEAU N°33 : Test DM à horizon de 4 périodes

	ARIMA	**ARX1**	**ARX2**	**ARX3**
ARIMA		0.9255	0.8128	0.9278
ARX1	0.07449		0.02411	0.08328
ARX2	0.1872	0.9759		0.9313
ARX3	0.07219	0.9167	0.06875	

Ici encore, aussi étrange que cela puisse paraître, les conclusions ne sont pas les mêmes que celles que nous avions pu émettre lors de la comparaison des prévisions sur les différents graphiques. D'après le tableau n°33 et au seuil de risque de 5%, le modèle ARX2 est moins précis que le modèle ARX1 - cela semble normal car il contient une variable de moins, donc sa qualité de prévision est inférieure. Si l'on considère un seuil de risque de 10% on voit que le modèle ARX3 a une capacité prédictionnelle supérieure aux modèles ARIMA et ARX2.

Le test DM étant basé sur les résidus des modèles et non les prévisions issues de ceux-ci, il semble en quelque sorte "normal" que le modèle composé de la production industrielle (ARX3) surpasse la capacité prévisionnelle du modèle composé de la popularité du mot 'emploi' (ARX2). Effectivement, son log de vraisemblance ainsi que son R^2 sont plus élevés, et son écart type est légèrement plus faible (confère estimation des modèles en section IV de cette partie). Au seuil de risque de 10% on voit aussi que le modèle ARX1 est meilleur que tous les autres modèles, à savoir ARIMA, ARX2 et ARX3.

TABLEAU N°34 : Test DM à horizon de 8 périodes

	ARIMA	**ARX1**	**ARX2**	**ARX3**
ARIMA		0.705	0.4251	0.605
ARX1	0.295		0.01749	0.07195
ARX2	0.5749	0.9825		0.9966
ARX3	0.395	0.928	0.003425	

Enfin, nous nous intéressons aux capacités de prédiction des modèles à un horizon de 8 périodes. Au seuil de risque auquel nous travaillons habituellement, on constate que le modèle ARX2 est moins précis que le modèle ARX1 comme c'était le cas aux horizons h=1 et h=4, il est aussi moins précis que le modèle ARX3. Ainsi, plus l'horizon devient grand, plus le modèle composé de X_3 devient précis d'un point de vue prévisionnel que le modèle composé de X_1.

De cette façon, à court terme le modèle ARX1 est plus précis que les modèles ARX2 et 3 qui n'incluent qu'une des 2 variables explicatives présentes dans l'ARX1, à moyen terme le modèle ARX1 surpasse le modèle ARX2 dans sa capacité prévisionnelle, et à long terme le modèle ARX2 est moins précis que les modèles ARX1 et 3. Aussi, le test de Dieblod-Mariano nous a montré que la capacité prévisionnelle du modèle incluant uniquement la variable Google Trends, est inférieure au modèle ARX1 aux 3 horizons différents, et inférieure au modèle ARX2 seulement à long terme (h=8). Nous pouvons alors penser que l'outil Google Trends est un bon indicateur des prévisions du taux de chômage lorsqu'il est combiné avec la production industrielle (comme c'est le cas dans le modèle ARX1). En revanche à long terme des indicateurs plus classiques sont davantage précis dans la prévision du taux de chômage.

Nous pouvons terminer cette étude en appliquant le test multiple de Mariano et Preve qui sélectionne le(s) modèle(s) ayant la meilleure capacité prédictive.

E) Tests multiples de Mariano et Preve

Le test multiple que contient la fonction **MDM.test** sous R ne permet pas de trouver quels modèles ont des capacités prévisionnelles exceptionnelles à un horizon de prévision d'une seule période. Nous ne l'appliquerons donc que pour comparer les prévisions à 4 et 8 périodes. Nous rappelons les correspondances des noms des prévisions :

- prev1 : prévisions réalisées à partir du modèle AR(1)
- prev2 : à partir du modèle ARX1 (avec X_1 la popularité du mot 'emploi' sous google Trends et X_3 la production industrielle)
- prev3 : à partir du modèle ARX2 (avec X_1)
- prev4 : à partir du modèle ARX3 (avec X_3)

À chacun de ces noms s'ajoute l'horizon de prévision tel que H1, H4 ou H8.

TABLEAU N°35 : Test de Mariano et Preve à horizon h=4

```
####################################################
Models with outstanding predictive ability:

         Rank       Sc Mean loss
prev1H4     3   3.1491    0.2738
prev2H4     2  -1.9623    0.0698
prev3H4     1 -16.0725    0.0873

p-value: 0.9836

Number of eliminated models:  1
####################################################
```

Mis en place sur les prévisions à horizon de 4 périodes, le test de Mariano et Preve dévoile que les modèles ARIMA, ARX1 et ARX2 disposent d'une capacité prévisionnelle exceptionnelle, puisque la p-value associée au test s'élève à 0,9836. En revanche, le modèle ARX3 a été éliminé, cela signifie que sa capacité de prévision est moins bonne que celle des autres modèles. Le test MDM classe également les modèles selon leur capacité prévisionnelle : on retrouve le modèle ARX2 en première place, suivit par l'ARX1 puis l'AR(1). En outre, d'après ce test c'est le modèle qui contient la variable de Google Tendances qui prédit au mieux le taux de chômage à moyen terme. Voyons ce que révèle le test appliqué sur les prévisions à h=8.

TABLEAU N°36 : Test de Mariano et Preve à horizon h=8

```
####################################################
Models with outstanding predictive ability:

         Rank       Sc Mean loss
prev1H8     2  2.2798    0.7389
prev2H8     1 -3.0267    0.2339
prev4H8                  0.3091

p-value: 0.9918

Number of eliminated models:  1
####################################################
```

Sur les prévisions à long terme on voit que le modèle ARX2 a été éliminé de la sélection de ceux ayant une capacité prévisionnelle supérieure, les 3 prévisions restantes ont la même précision puisque la valeur de la probabilité associée au test est 0,9918. De même, dans le classement on retrouve le modèle ARX1 (duquel sont issues les prévisions prev2H8) en première position, *i.e.* avec la meilleure capacité prédictive, puis le modèle ARIMA (prev1H8).

Il paraît très étrange que les prévisions du modèle AR(1) soient considérées aussi précises que celles des modèles ARX1 et 3 puisque comme nous l'avons remarqué et expliqué précédemment, les prévisions ARIMA à h=8 ne suivaient pas du tout les variations des valeurs observées du taux de chômage.

Finalement, étant donné que les différents critères que nous avons regardés pour comparer nos prévisions donnent des indications presque contradictoires, nous allons classer les 4 modèles aux 3 horizons de prévision selon le critère de comparaison, où 1 correspond au modèle ayant les meilleures prévisions et 4 celui ayant les prévisions les plus éloignées des vraies valeurs de la série. Lorsque la conclusion est la même aux 3 horizons nous ne mettons qu'un seul rang, sinon, nous distinguons selon les horizons de prévisions. Pour le test de Diebold et Mariano par exemple, nous effectuons le classement selon la p-value associée au test - nous regardons si pour un modèle elle est plus proche de 0, ou plus proche du seuil de significativité etc.

TABLEAU N°37 : Récapitulatif et classement des modèles

	Graphiques	**Taux d'erreur**	**Test DM**	**Test multiple**
ARIMA	4è	4è	- h=1 : 2è - h=4 : 3è - h=8 : 2è	- h=4 : 3è - h=8 : 2è
ARX1	1er	1er	1er	- h=4 : 2è - h=8 : 1er
ARX2	2è	2è	4è	- h=4 : 1er - h=8 : 4è
ARX3	3è	3è	- h=1 : 3è - h=4 : 2è - h=8 : 3è	- h=4 : 4è - h=8 : 3è

Conclusion & Discussion

De plus en plus présent dans notre société, le Big Data est devenu un facteur clé pour obtenir des informations pouvant aider l'analyse économique. Ainsi, dans cette étude nous avons essayé de prévoir le chômage et l'emploi à l'aide de l'outil Google Trends ainsi que d'autres variables économiques. La difficulté de l'utilisation de tels outils est de trouver quel mot représente le mieux la réalité des recherches internet des chercheurs d'emplois, nous avons donc regardé l'évolution de la popularité relative de 3 termes différents et avons choisi le mot 'emploi' comme variable prédictive du taux de chômage en France de 2004 à 2019.

L'utilisation de 'Google Trends' (GT) sur notre échantillon est concluante, sa cointégration avec Y_t révèle que l'outil peut aider à prévoir le chômage et l'emploi en étant inséré comme variable explicative sur des modèles auto régressifs incluant des variables exogènes (ARX). En effet, le modèle combinant une variable macroéconomique qu'est la production industrielle, couramment utilisée dans la prédiction du taux de chômage, à la popularité du mot 'emploi' sous google Trends, surpasse tout autre modèle en termes de prévision. Il est plus précis qu'un modèle n'incluant pas de variables explicatives, mais aussi qu'un modèle ne prenant pas en compte ce nouvel outil internet.

Les taux d'erreur sont croissants avec l'horizon de prévision, et nous n'avons pas noté d'impact particulier de l'horizon des prévisions quant à l'apport de Google trends pour prédire le chômage, excepté avec le test multiple de Mariano et Preve qui plaçait le modèle composé uniquement de la variable Google Tendances, en première position à h=4 et en dernière à h=8. En outre, l'outil semble plus efficace pour des prévisions de court ou moyen terme que pour du long terme, et est surtout efficace lorsqu'il est combiné aux variables macroéconomiques traditionnelles - comme nous l'avons expérimenté dans le modèle ARX1. Ainsi, malgré ses avantages incontournables car disponible gratuitement et en temps réel, nous nous intéressons aux limites intrinsèques à l'outil qui peuvent expliquer son apport décroissant avec l'horizon de prévision.

L'outil du géant Google est disponible en temps réel et est donc pertinent pour compléter les indicateurs économiques traditionnels. Mais les données qu'il fournit sont brutes, elles ne sont corrigées d'aucun effet saisonnier et ne sont ni stabilisées ni débarrassées d'éventuelles valeurs atypiques. Aussi, avant de cointégrer cette variable au taux de chômage, nous avons du la désaisonnaliser pour pouvoir l'utiliser correctement. La pérennité des

résultats de cet outil est encore trop fragile, c'est pourquoi il faut savoir s'en servir avec prudence. De plus, bien que les recherches Google représentent 95% des recherches françaises et 90%[45] des recherches mondiales, il y a un manque à gagner lié au développement des applications. En effet, les internautes accèdent de plus en plus à leurs requêtes en ayant recours aux applications pour réaliser leurs achats par exemple (les marques commencent à développer leurs propres applications maintenant). Finalement, une hausse de la tendance de recherche n'est pas toujours la conséquence d'une hausse des recherches en volume mais à une migration de la population qui peut faire baisser le nombre d'utilisateurs et peut ainsi fausser les résultats de l'outil Google Trends.

Notre analyse comportait initialement 5 variables explicatives, 3 d'entre elles ont pu être cointégrées à Y_t, constituant ainsi des variables à inclure dans les modèles ARX. Le taux d'intérêt et la population active qui n'ont pas pu être cointégrés sont, dans la littérature, de bons indicateurs dans l'explication du taux de chômage. Quelles raisons économiques peuvent expliquer les conclusions de notre étude pour ces 2 variables ? La période étudiée s'étend de 2004 à nos jours, elle prend donc en compte la crise des Subprimes qui est venue bouleverser l'équilibre économique et donc créer des évolutions de grande variance qu'il est difficile d'analyser par la suite. Ainsi, l'atypicité de la période étudiée a entraîné des fluctuations inhabituelles des indicateurs, ce qui explique que certaines séries ne convergent pas.

Il pourrait donc être intéressant de construire des modèles avant et après crise pour voir si les conclusions sont différentes. De même, pour améliorer les comparaisons entre les prévisions il serait bon d'inclure davantage de variables explicatives pour mieux mesurer l'apport de GT par rapport à de multiples variables, sur les prévisions du chômage. En effet, malgré la cointégration de la variable 'effectifs dans l'industrie manufacturière' avec Y_t, la relation qui les liait en étant différenciées s'est révélée nulle. Nous n'avons donc pas pu inclure cette dernière et nous sommes retrouvés avec 2 variables exogènes.

Pour conclure cette étude, on peut dire que Google Trends est un bon outil pour prédire le taux de chômage français grâce à ses nombreux avantages, mais qui doit être amélioré pour une utilisation officielle comme indicateur économique à part entière, notamment à plus long terme.

[45] https://www.planetoscope.com/Internet-/1474-recherches-sur-google.html (consluté le 17/02/2020)

Bibliographie

Amin S., “INDUSTRIE : Industrialisation et formes de société”, *Universalis*.
https://www.universalis.fr/encyclopedie/industrie-industrialisation-et-formes-de-societe/1-les-etapes-techniques-de-l-industrialisation/

Andrieu O., “Google Insights for Search disponible en français”, *Abondance*, 19/09/2019.
https://www.abondance.com/20090819-10009-google-insights-for-search-disponible-en-francais.html

Anota M., “Les taux d’intérêt négatifs peuvent aggraver le chômage keynésien”, *Alternatives economiques*, 27/05/2018.
https://blogs.alternatives-economiques.fr/anota/2018/05/27/les-taux-d-interet-negatifs-peuvent-aggraver-le-chomage-keynesien

Banque de France, “Enquêtes de conjoncture”, mis à jour le 08/04/2020.
https://www.banque-france.fr/statistiques/conjoncture/enquetes-de-conjoncture

Banque Mondiale Données, “Industrie, valeur ajoutée (% du PIB)”, consultable en ligne :
https://donnees.banquemondiale.org/indicateur/NV.IND.TOTL.ZS?view=chart

Bateman F., “Chapitre 3 : Identification”, *Laboratoire d'Information Systèmes - Aix Marseille Université*.
http://francois.bateman.free.fr/HTML/filtrage-ident/chapitre%203/ident4/IDENT4.htm

Bergel-Hayat R., “La prise en compte de variables explicatives dans les modèles de séries temporelles:application à la demande de transport et au risque routier”, *Université Paris-Est*, 2008, p.45. Consultable en ligne :
https://tel.archives-ouvertes.fr/tel-00432051/document

Bourbonnais R., “Économétrie”, *DUNOD Editions*, 2018, pp.258-296.

Bremme L., “Définition : Qu’est-ce que le Big Data ?”, *Le Big Data*, 07/2016.

https://www.lebigdata.fr/definition-big-data

Brief Eco, “Les déterminants du chômage”, 31/10/2018.
https://www.brief.eco/a/2018/10/31/on-fait-le-point/les-determinants-du-chomage/

Charpentier A., “Cours de séries temporelles”, *ENSAE Paris Dauphine*, pp.6-15.
https://www.math.u-bordeaux.fr/~hzhang/m2/st/TS1.pdf

Etner J. et Le Maitre P., “L'impact du taux d'intérêt sur l'évolution simultanée du chômage et de l'épargne”, *Persée*, 1999, pp. 917-935.
https://www.persee.fr/doc/reco_0035-2764_1999_num_50_5_410125

Garbay A, “Les Français cherchent un emploi sur Internet mais le trouvent grâce à leur réseau”, *Le Figaro*, 17/01/2017.
https://www.lefigaro.fr/emploi/2017/01/17/09005-20170117ARTFIG00290-les-francais-cherchent-un-emploi-sur-internet-mais-le-trouvent-grace-a-leur-reseau.php

Google Trends consultable en ligne :
https://trends.google.fr/trends/?geo=FR

Hellier J., “Économie du travail”, *Research Gate*, 01/2018.
https://www.researchgate.net/publication/322540093_Economie_du_Travail_Master_1_Dossier

INSEE, *Enquête Emploi*, 13/02/2020, n°36.
https://www.insee.fr/fr/statistiques/4309346

INSEE, définitions consultables en ligne :
https://www.insee.fr/fr/metadonnees/definitions

Inwin - Digital Expert, “Un outil puissant et gratuit mais méconnu Google Trends”, 11/03/2019.
https://www.inwin.fr/blog/un-outil-puissant-et-gratuit-mais-meconnu-google-trends/

Journal du net, “Google Trends (ex Google Insight) : définition”, mis à jour le 09/01/2019.
https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203505-google-trends-ex-google-insight-definition/

Journal du Net, “Nombre d'internautes en France”, mis à jour le 14/02/2014.
https://www.journaldunet.com/ebusiness/le-net/1071394-nombre-d-internautes-en-france/

Lambert M., “Comment passer de l’idée à la stratégie grâce à Google Trends”, *La Tranchée*, 11/10/2018.
https://www.latranchee.com/comment-passer-de-lidee-a-la-strategie-grace-a-google-trends/

Le Figaro, “La BCE maintient ses taux directeurs au plus bas”, 12/12/2019.
https://www.lefigaro.fr/conjoncture/la-bce-maintient-ses-taux-directeurs-au-plus-bas-1-20191212

Levasseur S., “Vieillissement de la population”, *Revues de l’OCFE*, 2015, pp. 339-370.
https://www.cairn.info/revue-de-l-ofce-2015-6-page-339.htm#

L’Express, “Taux de chômage à 8% en 2007”, 06/03/2008.
https://lexpansion.lexpress.fr/actualite-economique/taux-de-chomage-a-8-en-2007_470864.html

Ministère du Travail, “Les demandeurs d’emploi en janvier 2016”, 24/02/2016.
https://travail-emploi.gouv.fr/archives/archives-presse/archives-communiques-de-presse/article/les-demandeurs-d-emploi-en-janvier-2016

Moyou E., “Part des ménages ayant un accès internet en France de 2006 à 2018”, Statista, 08/01/2020.
https://fr.statista.com/statistiques/509227/menage-francais-acces-internet/

OCDE consultable en ligne :
https://data.oecd.org/fr/

Ouest France, “Le taux de chômage au plus bas depuis 10 ans”, 27/12/2019.

https://www.ouest-france.fr/economie/emploi/chomage/le-taux-de-chomage-au-plus-bas-depuis-dix-ans-6671745

Sénat, “Le réseau de la Banque de France - B. LES ENQUÊTES DE CONJONCTURE”. https://www.senat.fr/rap/r02-254/r02-2549.html

Tavernier J-L., “Note de conjoncture”, *INSEE*, 03/2015, pp.43-56. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=9&ved=2ahUKEwi2nLu42ojnAhWJzYUKHUuRBlEQFjAIegQICRAB&url=https%3A%2F%2Fwww.insee.fr%2Ffr%2Fstatistiques%2Ffichier%2F1408926%2Fmars2015_d2.pdf&usg=AOvVaw0Kee7rCU0qt0_SnakRGtid

The R Graph Gallery, consultable en ligne : https://www.r-graph-gallery.com/index.html

Tsay R. S., “Analysis of financial times series”, *University of Chicago*, 2002, pp.28-53.

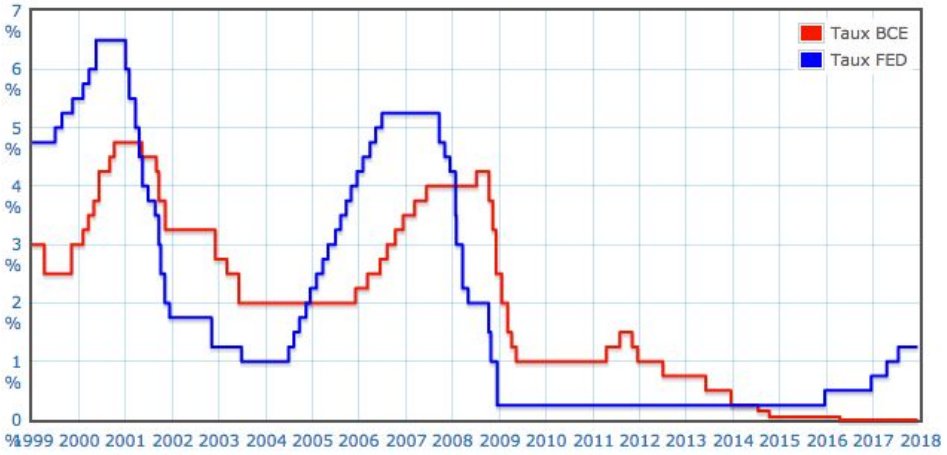
Vaté M., “Statistiques chronologiques et prévisions”, *Economica*, 1993, pp.217-225.

Vie publique, “Les grands secteurs de production : primaire, secondaire et tertiaire”, mis à jour le 03/09/2019. https://www.vie-publique.fr/fiches/269995-les-grands-secteurs-de-production-primaire-secondaire-et-tertiaire

Yassine A., “Google Trends : popularité ou volume de recherche d’un terme, de quoi parle-t-on ?”, *Ya-Graphic*, 04/07/2016. https://www.ya-graphic.com/google-trends-popularite-volume-de-recherche/

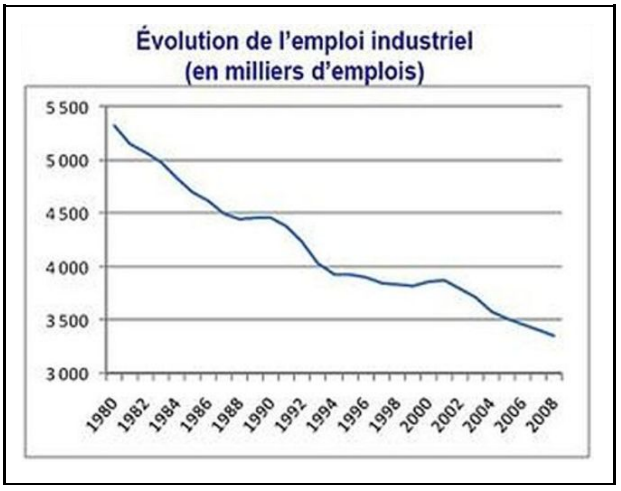
Annexes

Annexe n°1 : Évolution des taux directeurs de la FED et de la BCE depuis janvier 1999



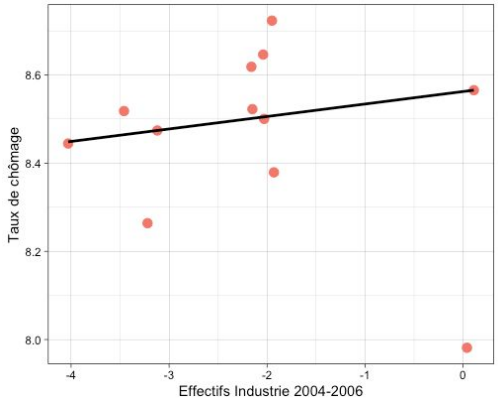
Source : https://france-inflation.com/taux-directeurs-bce-fed.php (consulté le 07/02/2020

Annexe n°2 : Évolution de l'emploi industriel de 1980 à 2008

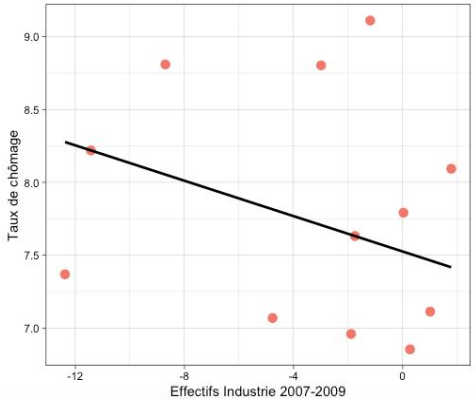


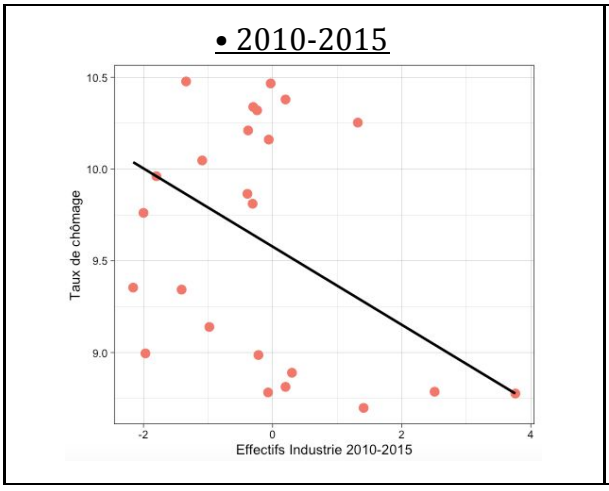
Annexe n°3 : Graphique de corrélation entre Y_t et X_5 selon les périodes

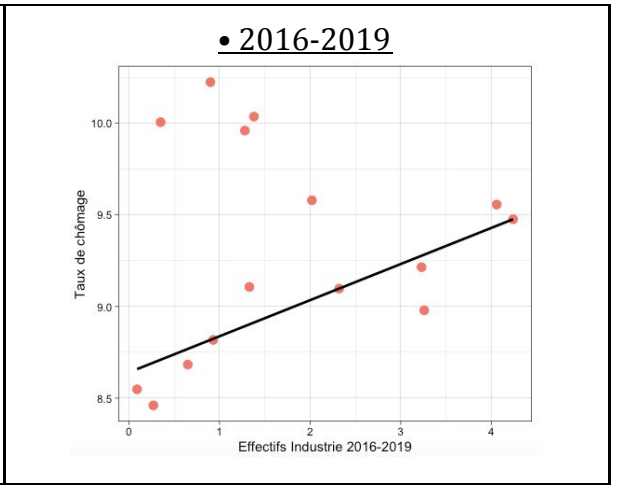
• 2004-2006



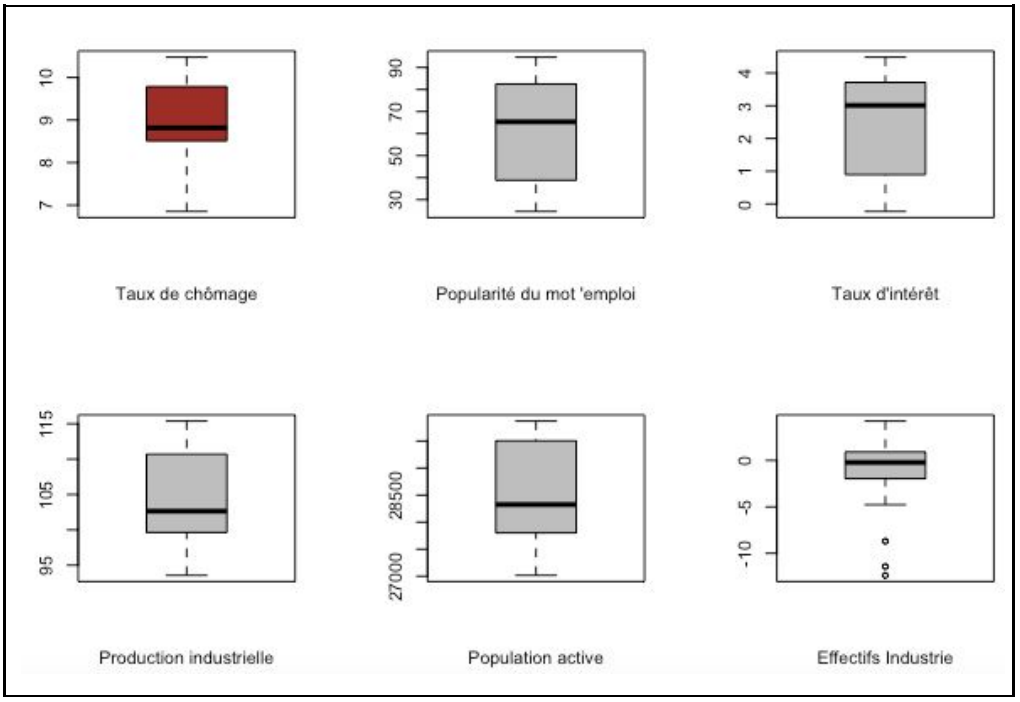
• 2007-2009



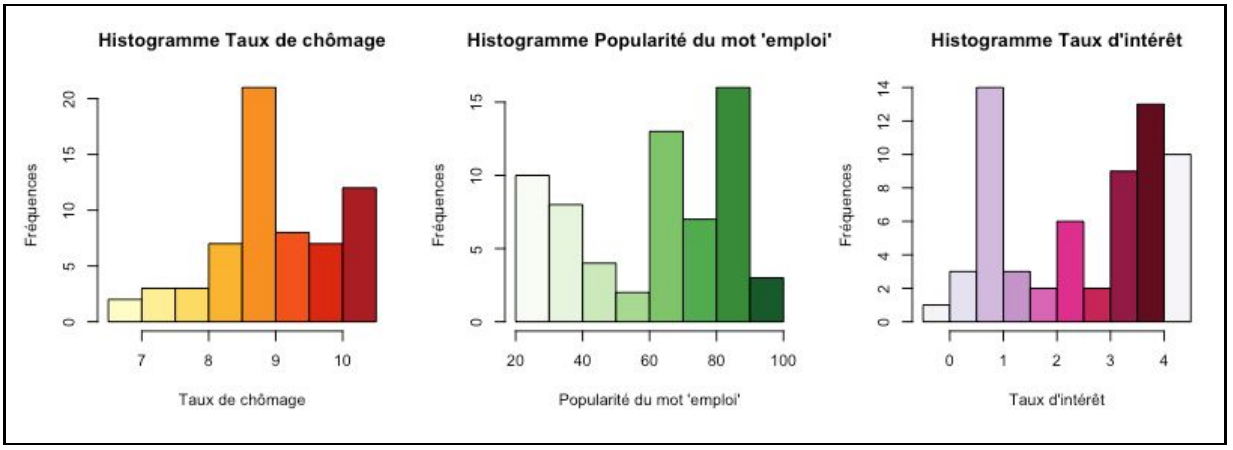


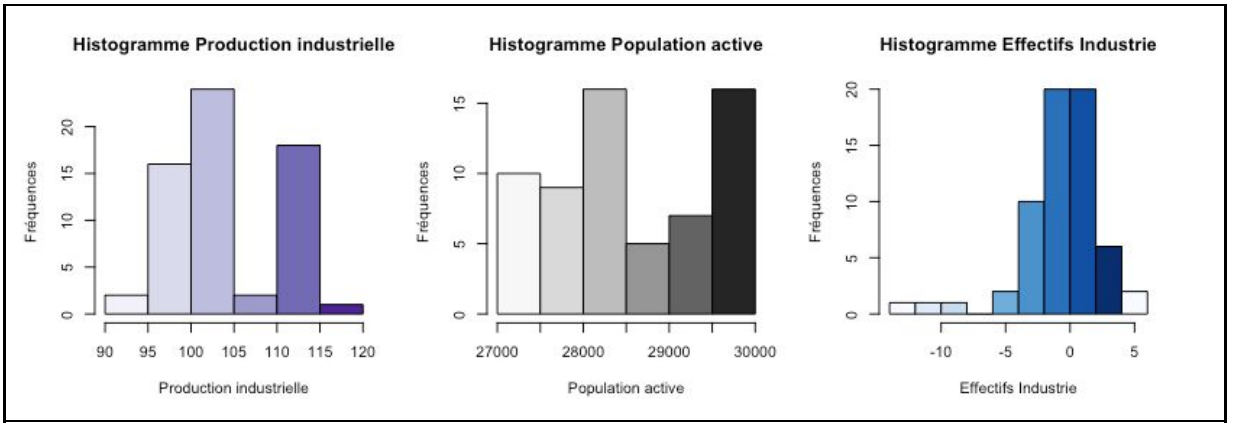


Annexe n°4 : Boxplots des 5 variables réalisés sous R studio



Annexe n°5 : Histogrammes de distribution des 6 variables



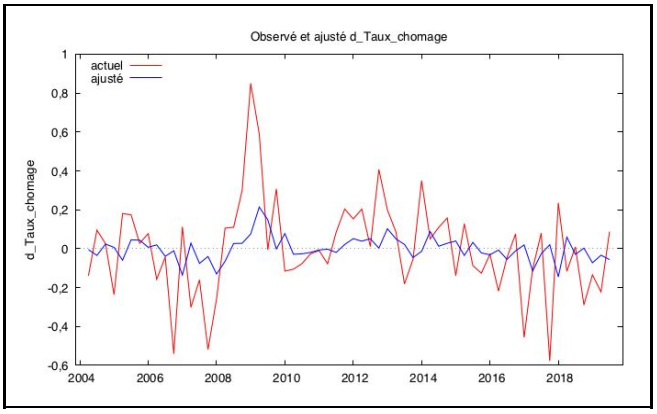


Annexe n°6 : Fonctions d'auto-corrélation de Yt différencié une fois

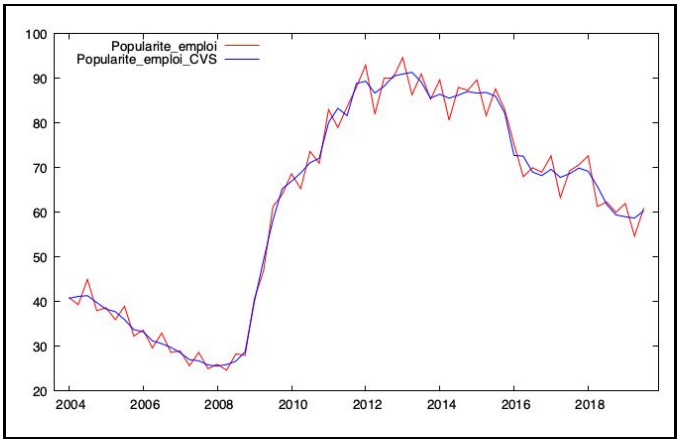
Fonction d'auto-corrélation pour d_Taux_chomage
***, **, * indicate significance at the 1%, 5%, 10% levels
using standard error 1/T^0,5

RETARD	ACF	PACF	Q	[p. crit.]
1	**0,2537** **	**0,2537** **	4,1855	[0,041]
2	**0,2916** **	**0,2428** *	9,8070	[0,007]
3	0,1770	0,0674	11,9131	[0,008]

Annexe n°7 : Graphique des valeurs prédites et des valeurs observées du modèle AR[1]



Annexe n°8 : Séries initiale et CVS de la variable "popularité du mot emploi"



Annexe n°9 : Sortie du modèle OLS de long terme entre le taux de chômage et le taux d'intérêt, et test KPSS des résidus

```
Modèle 7: MCO, utilisant les observations 2004:1-2019:3 (T = 63)
Variable dépendante: Taux_chomage

                coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------------
  const          10,0980      0,168045      60,09          5,42e-56   ***
  Taux_interet   -0,446613    0,0583922     -7,648         1,77e-10   ***

Moy. var. dép.          8,981320   Éc. type var. dép.      0,916827
Somme carrés résidus    26,60301   Éc. type de régression  0,660390
R2                      0,489537   R2 ajusté               0,481169
F(1, 61)                58,49945   p. critique (F)         1,77e-10
Log de vraisemblance   -62,23665   Critère d'Akaike        128,4733
Critère de Schwarz      132,7596   Hannan-Quinn            130,1591
rho                     0,967973   Durbin-Watson           0,157724
```

```
Test KPSS pour uhat7

T = 63
Paramètre du délai de troncation = 3
Statistique de test = 0,232453

                          10%      5%      1%
Valeurs critiques: 0,351   0,462   0,728
P. critique > .10
```

Annexe n°10 : Développement démarche de cointégration de Y_t avec X_3

```
Test de Dickey-Fuller augmenté pour Production_indus_manu
testing down from 20 lags, criterion AIC
taille de l'échantillon 54
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 8 retards de (1-L)Production_indus_manu
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,00162168
  statistique de test: tau_nc(1) = -0,650521
  p. critique asymptotique 0,4357
  Coeff. d'autocorrélation du 1er ordre pour e: 0,034
  différences retardées: F(8, 45) = 2,310 [0,0361]
```

```
Test de Dickey-Fuller augmenté pour d_Production_indus_
testing down from 20 lags, criterion AIC
taille de l'échantillon 54
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 7 retards de (1-L)d_Production_indus_manu
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -1,06309
  statistique de test: tau_nc(1) = -3,62015
  p. critique asymptotique 0,0002915
  Coeff. d'autocorrélation du 1er ordre pour e: 0,036
  différences retardées: F(7, 46) = 1,370 [0,2407]
```

```
Modèle 9: MCO, utilisant les observations 2004:1-2019:3 (T = 63)
Variable dépendante: Taux_chomage

                      coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------------------
  const                20,0043      1,48879       13,44          6,91e-20   ***
  Production_indus~    -0,105857    0,0142742     -7,416         4,46e-10   ***

Moy. var. dép.          8,981320   Éc. type var. dép.      0,916827
Somme carrés résidus    27,40643   Éc. type de régression  0,670288
R2                      0,474121   R2 ajusté               0,465500
F(1, 61)                54,99632   p. critique (F)         4,46e-10
Log de vraisemblance   -63,17388   Critère d'Akaike        130,3478
Critère de Schwarz      134,6340   Hannan-Quinn            132,0336
rho                     0,944581   Durbin-Watson           0,116061
```

```
Test KPSS pour uhat9

T = 63
Paramètre du délai de troncation = 3
Statistique de test = 0,383945

                          10%      5%      1%
Valeurs critiques: 0,351   0,462   0,728
P. critique interpolée 0,085
```

```
Modèle 11: MCO, utilisant les observations 2004:2-2019:3 (T = 62)
Variable dépendante: d_Taux_chomage

                      coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------------------
  const                -0,00445474   0,0261143     -0,1706        0,8651
  d_Production_ind~    -0,0508457    0,0144891     -3,509         0,0009     ***
  uhat9_1              -0,108863     0,0418107     -2,604         0,0116     **

Moy. var. dép.          0,000482   Éc. type var. dép.      0,245735
Somme carrés résidus    2,483029   Éc. type de régression  0,205147
R2                      0,325910   R2 ajusté               0,303059
F(2, 59)                14,26268   p. critique (F)         8,85e-06
Log de vraisemblance    11,77312   Critère d'Akaike       -17,54624
Critère de Schwarz     -11,16484   Hannan-Quinn           -15,04074
rho                     0,134515   Durbin-Watson           1,727062
```

Une différenciation a été nécessaire pour rendre X_3 stationnaire. Après estimation de la relation de long terme qui lie les 2 variables via une régression linéaire simple, on trouve que les résidus ceux eux aussi stationnaires. Enfin, lors de l'estimation de court terme on voit que la relation est pertinente car $\hat{\delta}$ est significatif et négatif.

Annexe n°11 : Développement démarche de cointégration de Y_t avec X_5

```
Test KPSS pour Effectifs_EMC

T = 63
Paramètre du délai de troncation = 3
Statistique de test = 0,590664

                          10%      5%      1%
Valeurs critiques: 0,351   0,462   0,728
P. critique interpolée 0,031
```

```
Test KPSS pour d_Effectifs_EMC

T = 62
Paramètre du délai de troncation = 3
Statistique de test = 0,034892

                          10%      5%      1%
Valeurs critiques: 0,351   0,462   0,728
P. critique > .10
```

```
Modèle 5: MCO, utilisant les observations 2004:1-2019:3 (T = 63)
Variable dépendante: Taux_chomage

                 coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------------
  Effectifs_EMC   -0,610569     0,364495       -1,675        0,0989     *

Moy. var. dép.          8,981320   Éc. type var. dép.       0,916827
Somme carrés résidus    4911,662   Éc. type de régression   8,900582
R2 non-centré           0,043298   R2 centré               -93,245725
F(1, 62)                2,805993   p. critique (F)          0,098950
Log de vraisemblance   -226,6145   Critère d'Akaike         455,2289
Critère de Schwarz      457,3721   Hannan-Quinn             456,0718
rho                     0,995796   Durbin-Watson            0,014996
```

```
Test KPSS pour uhat5(avec tendance)

T = 63
Paramètre du délai de troncation = 3
Statistique de test = 0,0842434

                          10%       5%       1%
Valeurs critiques: 0,121   0,148    0,214
P. critique > .10
```

```
Modèle 7: MCO, utilisant les observations 2004:2-2019:3 (T = 62)
Variable dépendante: d_Taux_chomage

                  coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------------
  const            0,441540     0,107545        4,106        0,0001     ***
  d_Effectifs_EMC -0,0190345    0,0158788      -1,199        0,2354
  uhat5_1         -0,0515933    0,0121689      -4,240        7,99e-05   ***

Moy. var. dép.          0,000482   Éc. type var. dép.       0,245735
Somme carrés résidus    2,811185   Éc. type de régression   0,218282
R2                      0,236822   R2 ajusté                0,210952
F(2, 59)                9,154159   p. critique (F)          0,000345
Log de vraisemblance    7,925184   Critère d'Akaike        -9,850368
Critère de Schwarz     -3,468965   Hannan-Quinn            -7,344868
rho                     0,096960   Durbin-Watson            1,786500
```

Il en va de même pour X_5 : elle est intégrée d'ordre 1 et ses résidus de long terme sont stationnaires. Lorsque l'on estima la relation de court terme qui la lie au taux de chômage, on s'aperçoit que delta est négatif et significatif : le relation n'est pas fallacieuse, les effectifs dans l'industrie sont cointégrés au taux de chômage.

Annexe n°12 : Estimation du modèle ARX2

```
> ARX2_1

Date: Mon May 25 10:59:30 2020
Dependent var.: data.matrix(Y)
Method: Ordinary Least Squares (OLS)
Variance-Covariance: Ordinary
No. of observations (mean eq.): 61
Sample: 2 to 62

Mean equation:

                     coef  std.error  t-stat  p-value
mconst          -0.0064038  0.0288744 -0.2218 0.825265
ar1              0.0881932  0.1281564  0.6882 0.494090
GTRENDS_CVS_diff 0.0294082  0.0092602  3.1758 0.002394

Diagnostics and fit:

                 Chi-sq df p-value
Ljung-Box AR(2)   1.4692  2 0.47970
Ljung-Box ARCH(1) 0.1343  1 0.71401

SE of regression 0.22432
R-squared        0.20338
Log-lik.(n=61)   6.11943
```

Annexe n°13 : Prévisions du taux de chômage à partir des modèles ARX à 3 horizons

→ Prévisions ARX à court terme

	ARX1	ARX2	ARX3
T3 2019	0.05777043	0.01998772	-1.545503e-06

→ Prévisions ARX à moyen terme

	ARX1	ARX2	ARX3

T4 2018	-0.137374378	-0.12694321	-0.053971413
T1 2019	-0.045050487	-0.06385308	0.016501617
T2 2019	-0.106957861	-0.03757221	-0.138391445
T3 2019	-0.004416715	-0.02421454	-0.007281679

→ Prévisions ARX à long terme

	ARX1	**ARX2**	**ARX3**
T4 2017	-0.046275666	0.009052287	-0.078046911
T1 2018	-0.047836915	0.052820965	-0.066409730
T2 2018	0.019727645	-0.116264561	-0.026048912
T3 2018	-0.066685267	-0.078448526	0.054520452
T4 2018	-0.137374378	-0.126943214	-0.053971413
T1 2019	-0.045050487	-0.063853078	0.016501617
T2 2019	-0.106957861	-0.037572215	-0.138391445
T3 2019	-0.004416715	-0.024214535	-0.007281679

Annexe n°14 : Tableau des erreurs de prévisions des 4 modèles

```
> df_erreur
   modeles erreur     H.1     H.4     H.8
1    ARIMA    MSE 0.01644 0.27375 0.73895
2     ARX1    MSE 0.00091 0.06980 0.23391
3     ARX2    MSE 0.00461 0.08734 0.27863
4     ARX3    MSE 0.00772 0.14184 0.30912
5    ARIMA    MAE 0.12820 0.50288 0.80544
6     ARX1    MAE 0.03009 0.25389 0.46496
7     ARX2    MAE 0.06787 0.27997 0.51205
8     ARX3    MAE 0.08786 0.36703 0.51492
9    ARIMA   MDSE 0.01644 0.27545 0.56203
10    ARX1   MDSE 0.00091 0.06424 0.24682
11    ARX2   MDSE 0.00461 0.07403 0.26573
12    ARX3   MDSE 0.00772 0.14519 0.28628
13   ARIMA   MDAE 0.12820 0.51919 0.73955
14    ARX1   MDAE 0.03009 0.25319 0.49556
15    ARX2   MDAE 0.06787 0.26961 0.51428
16    ARX3   MDAE 0.08786 0.38100 0.53392
```

Table des matières

Sommaire **1**

Résumé **2**

Introduction **3**

Partie 1 : Environnement économique **7**

I. Analyse économique du sujet 7

II. Analyse des variables explicatives 10

Google Trends : X_1 10

Taux d'intérêt : X_2 13

Production industrielle : X_3 16

Population active : X_4 17

Évolution des effectifs dans l'industrie manufacturière : X_5 19

Partie 2 : Méthodologie économétrique **23**

I. Modélisations et prévisions ARIMA 23

1- Stationnarisation de la série 25

2- Identification du modèle 26

3- Estimation du modèle 27

4- Vérification des résidus 28

5- Prévisions 29

II. Cointégration des variables 30

III. Modélisations et prévisions ARX 32

IV. Comparaison des prévisions 34

Partie 3 : Présentation des données et application **36**

I. Présentation des données 36

II. Modélisations et prévisions ARIMA 42

Stationnarisation de la série 42

Identification et estimation du modèle 44

Vérification des résidus 46

Prévisions 48

III. Cointégration des variables 50

Test de saisonnalité 50

Cointégration de la popularité du mot 'emploi' (X_1) 52

Cointégration du taux d'intérêt (X_2) 55

Cointégration de la production industrielle (X_3) 56

Cointégration de la population active (X_4) 56

Cointégration des effectifs dans l'industrie (X_5) 58
IV. Modélisations et prévisions ARX 59
Modélisations 60
Prévisions 64
V. Comparaison des prévisions 64
Transformation des prévisions 64
Représentations graphiques des prévisions 65
Erreurs de prévisions 67
Tests de Diebold-Mariano 69
Tests multiples de Mariano et Preve 71

Conclusion & Discussion 74

Bibliographie 76

Annexes 80

Annexe n°1 : Évolution des taux directeurs de la FED et de la BCE depuis janvier 1999 80
Annexe n°2 : Évolution de l'emploi industriel de 1980 à 2008 80
Annexe n°3 : Graphique de corrélation entre Yt et X_5 selon les périodes 80
Annexe n°4 : Boxplots des 5 variables réalisés sous R studio 81
Annexe n°5 : Histogrammes de distribution des 6 variables 81
Annexe n°6 : Fonctions d'auto-corrélation de Yt différencié une fois 82
Annexe n°7 : Graphique des valeurs prédites et des valeurs observées du modèle AR[1] 82
Annexe n°8 : Séries initiale et CVS de la variable "popularité du mot emploi" 82
Annexe n°9 : Sortie du modèle OLS de long terme entre le taux de chômage et le taux d'intérêt, et test KPSS des résidus 83
Annexe n°10 : Développement démarche de cointégration de Yt avec X_3 83
Annexe n°11 : Développement démarche de cointégration de Yt avec X_5 83
Annexe n°12 : Estimation du modèle ARX2 84
Annexe n°13 : Prévisions du taux de chômage à partir des modèles ARX à 3 horizons 84
Annexe n°14 : Tableau des erreurs de prévisions des 4 modèles 85